University of Nevada, Reno

Using the ROC Based Gini Coefficient to Quantify Spatio-Temporal Clustering of Earthquakes

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in
Statistics and Data Science

by

Natalie T. Bladis

Dr. Ilya Zaliapin/Thesis Advisor

May, 2023

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

NATALIE TEVA BLADIS

entitled

Using the ROC Based Gini Coefficient to Quantify Spatio-Temporal Clustering of Earthquakes

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Mihye Ahn, Ph.D.
Advisor

Deena Schmidt, Ph.D.
Committee Member

Daniel Trugman, Ph.D.
Graduate School Representative

Markus Kemmelmeier, Ph.D., Dean
Graduate School

May, 2023

Abstract

The purpose of this thesis is to explore a method for measuring the correlation between two measures or random variables using the ROC (Receiver Operating Characteristic) diagram and Gini coefficient. This procedure is then applied to seismic data to produce a metric of earthquake clustering in various regions.

Earthquake clustering is a fundamental component of seismicity that reflects various forms of earthquake triggering mechanisms. Zaliapin and Ben-Zion (2021) introduced a simple and robust measure of space-time clustering, using the ROC diagram, that allows disentangling effects related to concentration of events around a heterogeneous regional fault network (marginal space distribution of events) from coupled space-time fluctuations (joint space-time distribution). This work describes the mathematical and statistical foundation of their approach.

Specifically, this study:

- examines and illustrates seismic clustering in multiple seismically active regions, including the Reno area,
- explores several general measures of seismic rate that can account for the number of events, the total area of faultbreaks, seismic moment, and more,
- And systematically examines general and coupled space-time clustering of raw and declustered catalogs.

Conclusion of this analysis are that the overall observed earthquake clustering is high, for a variety of regional catalogs and global seismicity. At the same time, when the marginal clustering is removed, different catalogs show different degrees of coupled space-time clustering, reflecting a variety of specific triggering conditions and mechanisms.

Keywords: ROC Diagram, Gini Coefficient, Seismology.

Acknowledgment

I would like to express my sincere gratitude to my thesis advisor, Dr. Ilya Zaliapin, for his guidance, support, and invaluable feedback throughout my research. His knowledge and expertise have been instrumental in shaping my understanding of statistics and in helping me to develop as a researcher. When I came to him seeking a thesis advisor during my first semester, I had a tall list of things I wanted to accomplish and experiences while in graduate school, from presenting at conferences to obtaining funding to alleviate the financial burden of attend school to doing valuable research that impacts our global community. Dr. Zaliapin helped me achieve everything I wanted to during my graduate school experience. He was the kindest, most compassionate mentor. I hope my thesis can be a small contribution to his legacy.

I would also like to thank the members of my thesis committee, Dr. Deena Schmidt, Dr. Mihye Ahn, and Dr. Daniel Trugman, for their time and insightful comments on my work. Their constructive feedback and suggestions have greatly improved the quality of my research.

I would like to acknowledge the support and encouragement of my family and friends, who have always been there for me throughout my academic journey. Their unwavering belief in me has been a constant source of motivation and inspiration. I would like to give a big shout out to my cat, Echo, for all her unconditional love.

Thank you all for your invaluable contributions to my thesis.

Contents

List of Tables

List of Figures

1 Introduction

Forecasting extreme events is a fundamental task for the statistics community. A first step towards this goal is developing methods for finding relationships between several processes. This work proposes one such method, using the ROC(Receiver Operating Characteristic) and the Gini coefficient(G). A ROC diagram is a useful tool to compare measures on the same domain and to quantify the degree of concentration of a given measure. This method is applied to seismic data from earthquakes, which are extreme events that pose a significant threat to society and infrastructure. Seismology is a field of study that is concerned with the analysis of earthquakes, their behavior, and the impact they have on the environment. Statistical seismology is a subfield of seismology that aims to analyze earthquake data using mathematical and statistical methods. In recent years, the availability of large earthquake catalogs, such as those provided by the Advanced National Seismic System (ANSS), has provided an abundance of data for studying the behavior of earthquakes.

Understanding the complex seismic systems and then developing models that can be used for forecasting is a big task. This work focuses on quantifying one element of seismicity, earthquake clustering. To establish a foundation for the basis of this work, statistical models of seismicity are discusses in the literature review section, with a focus on methods that quantify clustering. Classic models, like the ETAS model, are presented along with models that try to characterize seismicity through a localization approach. Earthquake clustering is discussed as a key component of any predictive model for seismicity. Current methods of quantifying clustering and previous studies of clustering in various regions are presented, which leads us to the motivation of this work.

In the background section we give an overview of different mathematical and statistical approaches to quantifying correlation. The methodology section presents a mathematical foundation for determining how concentrated one measure is. A method for comparing a given measure first to a uniform measure, then to a more complex measure,

using the Receiver Operating Characteristic (ROC) diagram and the Gini Coefficient is discussed.

This work applies these techniques to earthquake data from the ANSS catalogs for the regions Reno, southern California, Japan, New Zealand, Italy, the Atlantic, and the Pacific. The Gini coefficient and ROC diagrams are used to quantify, characterize and compare clustering and the spatial-temporal distribution of seismic activity in the study regions.

The findings of this study will contribute to the existing knowledge of statistical seismology, provide insights into the spatial and temporal distribution of seismic activity in the regions studied, and present the statistical and mathematically theory behind these approaches. This research has implications for seismic hazard assessment and earthquake forecasting, which can inform policies and practices for disaster risk reduction and management.

2 Literature Review

2.1 Earthquake Catalogs

Earthquake catalogs from different regions around the globe provide a record of seismic activity reporting at-least the epicenter of each shock, origin time and magnitude. We will define an earthquake catalog as follows:

$$\mathbb{C} = \{t_i, m_i, \theta_i, \lambda_i, z_i\} \tag{1}$$

where t_i is the time, m_i is the magnitude , θ_i is the latitude, λ_i is the longitude and z_i is the depth of the ith event.

In this work, the following catalogs will be referred to:

- Northern California Earthquake Data Center (NCEDC), global catalog, https://ncedc.org/ncedc/catalog-search.html

- Hauksson et al. (2012), relocated catalog for southern California, https://scedc.caltech.edu/data/alt-2011-dd-hauksson-yang-shearer.html

- Hauksson et al. (2013), waveform-relocated Southern California earthquake catalog, https://scedc.caltech.edu/ data/alt-2011-dd-hauksson-yang-shearer.html

- Richards-Dinger and Shearer (2000), southern California, https://scedc.caltech.edu/data/alt-2000-richards.html

- ANSS catalog, global catalog, https://earthquake.usgs.gov/data/comcat/

2.2 Statistical Models of Seismicity

There are many models that describe seismic activity. Some of these models are based on the physical principles that govern earthquakes, while other models are statistical and developed from the catalogs of earthquake data. This thesis focuses on statistical models.

2.2.1 Point Process Models

Upon inspecting earthquake catalogs, Ogata et al. concluded that seismic activity in a region follows a time series with an extremely complicated structure. Seismic events can be thought of as a marked point process, where the origin times of each earthquake, t_i is modeled as a point process and all other characteristics are marks. Ogata et al.'s work analyzes a group of parametric models used in statistical analysis of earthquake catalogs and assesses earthquake risk in a region (Ogata, 1998).

The conditional intensity function plays an important role in the likelihood theory of point processes (Daley and Vere-Jones, 1972). The conditional intensity function can be understood roughly as the derivative of the probability of an event occurring at a time t

$$\lambda(t|F_t) = \lim_{\Delta \to 0} \mathsf{Prob}\left\{\text{An event occurs} \in (\mathsf{t}, \mathsf{t}+\Delta)/\mathsf{F_t}\right\}/\Delta, \tag{2}$$

where F_t is information over the time interval $(0, t)$ of observations available, including the history of the point process itself at time t.

2.2.2 Cyclicity Models

Two classes of models for interpreting seismic data are proposed by Ogata (1998). The first aims to break down seismicity into components of evolutionary trend, clustering and periodicity. To determine if sesimicity happens in cycles the uniformity of superposed point process is checked on one cycle (Shimazaki, 1971). These techniques for determining cyclicity assume the data is not contaminated by the presence of clustered events or by the change in the detection rate of seismicty that could uncover real or artificial trends (Vere-Jones and Ozaki, 1982; Vere-Jones, 1985). Data sets for analysis therefore are significantly decreased in size because of declustering and having a threshold for smaller events that ensures the homogeneous detection of sesimicity. Ogata provides a model to overcome these limitations.

Ogata (1983) suggested the following model for the conditional intensity in Equation (2),

$$\lambda(t|F_t) = a_0 + P_J(t) + C_K(t) + \sum_{t_i<t} g_M(t - t_i). \tag{3}$$

On the right-hand side of (3), the second term represents the evolutionary trend where

$$P_J(t) = \sum_{j=1}^{J} a_j \phi_j(t|T), \quad 0 < t < T, \tag{4}$$

T is the total length of the observed interval and $\phi_j(\cdot)$ is a polynomial of order j. These components are introduced to incorporate either the genuine seismic trend, the evolutionary change of the detection rate of shocks, or both. The third term of 3 is the Fourier expansion

$$C_K(t) = \sum_{k=1}^{K} \left\{ b_{2k-1} \cos(2k\pi t/T_0) \sin(2k\pi t/T_0) \right\}, \tag{5}$$

for cyclic effects with a given fixed cycle length T_0. The last term in (3) stands for the clustering effects such as aftershocks and earthquake swarms. The function $g_M(x)$ measures the increase in clustering due to a shock. This function is referred to as the *response function of a shock*, and parameterize it as

$$g_M(x) = \sum_{m=1}^{M} c_m x^{m-1} e^{-\alpha x}. \tag{6}$$

If in (6) the scaling parameter α is fixed, then the model in (3) is linearly parameterized. This model is then used to examine and establish the existence of each component by the comparison of AIC values among a possible set of (J, K, M) configurations. For example, if the configuration with $K = 0$ is selected, that is to say $C_0(t) = 0$, then this suggests that no periodicity of T_0 exists in the seismic actiivty, Otherwise,m its shape is estimated by the maximum likelihood method for the selected configuration of $(\hat{J}, \hat{K}, \hat{M})$.

2.2.3 Cluster Models

Another class of models is used to attempt to determine if earthquake sequences in two regions have a causal connection. The second-order properties between point processes such as cross correlation have been used to examine this problem (Brillinger, 1988). One main difficultly of this approach is that even if you detect a significant cross correlation between two realizations of point processes this method cannot discern which process causes the other, if they both cause each other to occur, or if another process entirely causes both. To make this distinction, Ogata and Akaike (1982) and Ogata et al. (1982) constructed a parametric model that applies the minimum AIC procedure using the mutually-exciting process by Hawkes' (Hawkes, 1971). To examine whether or not some other process causes the two correlation processes, call them $\{t_i\}$ and $\{u_i\}$, Ogata (1983b) extends the model in (3) to

$$\lambda(t|F_t) = a_0 + P_J(t) + C_K(t) + \sum_{t_i<t} g_M(t-t_i) + \sum_{u_j<t} h_N(t-u_j), \tag{7}$$

where $\{u_i\}$ is another series of events considered as inputs of the conditional intensity function. The response function $h_N(x)$ is a parameterized by

$$h_N(x) = \sum_{n=1}^{N} d_n X^{n-1} e^{-\beta x}. \tag{8}$$

It is expected that $h_N(x) = 0$ if there is no causal relation from $\{u_i\}$ to the conditional intensity function $\lambda_\theta(t|F_t)$, or the occurrence of $\{t_i\}$. If there is a causal relationship, then it would be important to determine how influential the relationship is and the approximate shape of the response function.

The correlation between the deep earthquakes beneath the Hida region in central Japan and the shallow earthquakes in central Kanto were explored using this method(Utsu, 1975). The outcomes of these analyses showed that Hida shocks can be classified as a stationary Poisson process, and that earthquakes in the central Kanto region receive significant one-way stimulation from earthquake occurrences in the Hida region and they are self-exciting (Ogata, 1998).

2.2.4 ETAS Model

Since the 1970s, stochastic branching processes were regularly utilized to model earthquake occurrence. Ogata (1988) developed the Epidemic Type Aftershock Sequence (ETAS) model, which combines the basic empirical principles of statistical seismology with rigorous stochastic modeling and estimate methods; see also Ogata (1985). The model represents seismic activity of earthquakes with magnitudes greater than M_0 regionally. A Poisson process with intensity $\mu(t)$ is used to model background events. A modified Omori law describes the offspring produced by each earthquake(Utsu, 1961).

The modified Omori law states that an earthquake with magnitude M_i that occurred at time t_i produces offspring according to a Poisson process with intensity

$$v(t|t_i, M_i) = \frac{K_0 10^{\alpha(M_i - M_0)}}{(t - t_i + c)^p}, \quad t > t_i \tag{9}$$

where the parameters are the positive constants K_0, α, c and $p > 1$. The numerator of the offspring intensity v is an exponential productivity law and the denominator is a power-law for temporal decay. According to the Gutenberg-Richter law,

$$P(M_i > M) = 10^{-b(M - M_0)}, \quad M > M_0. \tag{10}$$

each newly created event, background or offspring, is given a magnitude M_i independent of prior events, including its parent (Gutenberg & Richter, 1954). Background events, first-generation aftershocks, offspring of these aftershocks (called second generation aftershocks), offspring of the second generation aftershocks (third generation aftershocks), and so on form the combined earthquake flow (Kovchegov, Zaliapin, Ben-Zion, 2022).

The combined flow is a point process described by the ETAS model that will be discuss in detail in this section.

Conditional Intensity The normal seismic activity of a wide region is described by the conditional intensity function of the Epidemic Type Aftershock Sequence model (ETAS) (Ogata, 1998). This model superimposes a constant rate for background seismicity and the modified Omori functions of any shocks i which occurred at time t_i as follow:

$$\lambda(t|H_t) = \mu + \sum_{t_i < t} \frac{K_i}{(t - t_i + c)^p}, \tag{11}$$

where μ is an occurrence rate for the background seismic activity and $H_t = (T_i, M_i); t_i < t$ is the history of occurrence times t_i up to time t and their corresponding magnitudes M_i (Ogata, 1998). In this model the parameter K_i is defined as

$$K_i = K_0 e^{\alpha(M_i - M_0)} \tag{12}$$

and is dependent of the magnitude M_i of event i and the cut-off magnitude M_0 of the data set. Some advantages of the ETAS model compared to alternative models are that is does not require the discrimination between mainshocks and aftershock and it has the best fit when comparing goodness-of-fit. Among the parameters $\theta = (\mu, K_0, c, \alpha, p)$ of the ETAS model, the last two parameters α and p characterize the temporal pattern of seismicity. The α value measures magnitude sensitivity of an earthquake in generating its aftershocks in a wide sense and the p value represents the decay rate of aftershocks (Ogata, 1998).

Estimation The Omori formula quantifying the frequency of aftershocks per unit time interval is

$$n(t) = K(t + c)^{-p} \quad (K, c, p : parameters) \tag{13}$$

where t is the lapse time from the occurrence of the main shock, K depends on the magnitude of the main shock and the lower bound of the magnitude of aftershocks counted, and p is known to be independent of these(Utsu, 1961). If $p = 1$ this formula is called the original Omori formula and for $p \neq 1$ it is called the modified Omori formula. One interpretation of the value of p is that it represents physical state of the Earth's crust. For instance, there is a consistent regional fluctuation of the p value in Japan, related to variation in surface heat-flow values. Areas with greater crustal temperatures see a faster decrease of aftershock activity, or a faster release of stress (Mogi, 1962; Ogata, 1998).

Traditional estimates of the parameter p and the other three parameters of the modified Omori formula are outlined by Utsu (1961) and Ogata (1983a,b). For these estimates, we concider occurence time of afterschocks sequence $t_1, t_2, ..., t_N$ in a time interval S, T, where the occurence time of the main shock corresponds to the origin of

the time axis, $t = 0$. Using the modified Omori formula to represent the aftershock sequence, assume that the afterschock sequence is distributed according to a nonstationary Poisson Process with intensity function

$$\lambda(t;\theta) = K(t+c)^{-p}, \quad \theta = (K, c, p), \tag{14}$$

To estimate the parameters, maximizing the log-likelihood function of the aftershock sequence with respect to the parameters (K, c, p) to obtain the maximum likelihood estimate (MLE) $\hat{\theta} = (\hat{K}, \hat{c}, \hat{p})$. The inverse of the Fisher information matrix $J(\hat{\theta})^{-1}$ produces the variance-covariance matrix of the errors of the MLE. The modified Omori formula's maximum likelihood estimation makes it feasible to predict the probability of major aftershocks (Ogata, 1998).

In the calculation of the log-likelihood, a suitable starting time S must be determined to avoid significant bias and to gain accurate estimation of p and c. If the aftershocks were detected homogeneously throughout the entire observation period $[0, T]$, then we can formulate $S = 0$. However, in some cases the MLE computed with $S = 0$ will produce biased estimates, if the aftershock sequences are too complicated, for example. Methods for choosing S include selecting the S that produces the smallest AIC value or the magnitude versus time plots can be visually inspected to calculate S in a less exact manner (Ogata, 1998). The model (14) is inaccurate if an earthquake sequence over an observed time period does not consist of a single pure aftershock sequence.

Ogata proposes a transformation of the conditional intensity with respect to time to a frequency-linearized time so that the occurrence of earthquakes becomes the standard stationary Poisson process. He used this technique to show that the intensity trend of an earthquake's aftershocks decays over a lengthy period of time, if the background seismicity level in the vicinity of an aftershock zone is very low. The alternative intensity functions can also be used to model secondary aftershocks. Additionally, sometimes the rate of aftershock decay in some stage can occur more quickly than predicted according to the modified Omori formula. A shift from aftershock to normal activity may

be directly observed and objectively detected using the techniques and models Ogata describes. In order to recognize a seismic quiescence that could occur before the next large earthquake, it is essential to identify and quantify normal activity (Ogata, 1998).

The change-point problem can be used to determine whether or not the temporal patterns of seismicity changed before and after a time T_0 in a given data set on a time interval $[S, T]$. AIC values are computed as the criterion for model selection to determine if a model throughout the time interval $[S, T]$ fits better than two different models on the respective intervals $[S, T_0]$ and $[T_0, T]$ for some T_0. The method of selecting T_0 dictates the way AIC values are compared and computed. More specifically, if a possible change-point T_0 is chosen using information other than observed data, such as a theoretical reason, different techniques are used than when T_0 is determined from the data. Essentially, if the AIC value for the entire interval $[S, T]$ is larger than the sum of the AIC values for the interval $[S, T_0]$ and $[T_0, T]$, with some adjustments made when T_0 is estimated from the data itself, then there is a significant distinction in the patterns of seismicity in the two divided time spans. Otherwise, we deduce that there isn't a change-point there. To find more potential change-points in the range $[S, T_0]$, repeat this technique (Ogata, 1998).

Ogata describe how to use the point process equivalent of residual analysis for regression models to identify periods of relative quiescence. Origin times t_i are transformed 1-to-1 into the residual point process (RPP) τ_i. If the maximum likelihood estimate (MLE) $\hat{\theta} = (\hat{\mu}, \hat{K_0}, \hat{c}, \hat{\alpha}, \hat{p})$ of the ETAS model for a data set provides a good fit to the seismicity, then the RPP is approximately the standard stationary Poisson process. If a major departure of any characteristic property of the RPP from that anticipated for a stationary Poisson is observed, then there is a disparity between the model and data. Non-homogeneous data or the presence of seismic quiescence could cause such a deviation because these structures are not best modeled by the ETAS equations. For example, during a period of relative quiescence, the occurrence rate of the RPP will be significant smaller than that expected from the standard Poisson process. This more

mathematical approach was a significant improvement on the qualitative methods of reporting quiescence at the time. Identifying periods of relative quinescence is important for the prediction of future seismic activity. Understanding whether these periods of relative inactivity signal the end of aftershock activity or are the precursor of a forthcoming larger earthquake, for example, will be essential in the prediction process (Ogata, 1998).

The ETAS model is a higher-order approximation of the aftershock occurrences than the modified Omori formula. The modified Omori formula models aftershock sequences as simple inverse power decay, but in many cases aftershock sequences are more complex. The ETAS model better captures these complexities and fits well to various types of aftershock sequences (Ogata, 1998).

In the last section, Ogata provides an extension of the ETAS model to create a space-time conditional intensity function that models a history-dependent probability that an earthquake occurs. The specifics of choosing the parameters of this function to best model the space-time point process of seismicity are described in detail (Ogata, 1998).

By using a hypocenter database, point-process models described by parameterized conditional intensity functions can be effective instruments for analyzing seismic activity. AIC evaluates the model's goodness-of-fit and can help distinguish between competing hypotheses. The maximum likelihood estimation approach is used to estimate the modified Omori formula for aftershock decay effectively. The ETAS model quantifies the characteristics of a focus region's standard seismicity. Modeling the intensity function in respect to the crustal stress-field change will be necessary. Using data on anomalous occurrences like relative quiescence, point-process modeling aims to create an algorithm for the practical application of probability predictions of major earthquakes. Using spatial, time, and magnitude pattern, Ogata et al. (1995, 1996) attempted to develop real-time probability discrimination of foreshocks. It is suggested to assess probability forecasting using the AIC difference. Ultimately, point-process modeling could be applied to forecast the probability of large earthquakes (Ogata, 1998).

2.3 Localization Approach - Progressive Localization

The mechanisms generating massive earthquakes are still a mystery after decades of observational, experimental, and theoretical research (Ben-Zion, 2008). Kato and Ben-Zion (2021) synthesize existing knowledge of the starting mechanisms of major earthquakes and present some key characteristics of the origin processes. They discuss evidence of earthquake-induced rock damage causing regional weakening, and that a few years prior to certain large earthquakes there is progressive localization of deformation around the ensuing rupture zones. Depending on the circumstances, a combination of slow slip transients and foreshocks at various spatial and temporal scales comprise the last stage of deformation localization. In contrast to the smooth acceleration anticipated for a developing aseismic nucleation phase, the development of slip on large, localized faults displays a step-like rise. Ultimately, they propose an integrated model to explain the diversity of large earthquake generation that uses multiple models at different scales simultaneously to describe a seismic system.

Large earthquakes that pose substantial social and economic threats to people all over the world. Understanding the processes that lead up to large seismic events is a pivotal question for the seismology community (Abercrombie, 2019). Changes in the amount of stress and strength in a fault zone are the underlying elements that lead up to a large, dynamic rupture (Ben-Zion et al., 2003). This paper outlines three main models to describe the process- cascade-up, pre-slip and progressive localization.

Large earthquakes on a heterogeneous fault can be modeled by the cascade-up structure as forming in reaction to static and dynamic stress fluctuations brought on by earlier earthquakes, which locally increase the long-term tectonic stress (Ellsworth, 2018). Large mainshock ruptures are the result of a series of seismic events on pre-existing faults. Until the dynamic fracture ends, the earthquake's magnitude is uncertain.

The pre-slip model, in contrast, describes the generation of significant events on a fault surface that is rather uniform. These are started by gradual slip or fluid movement and may cause later, more powerful earthquakes (Dieterich, 1992). The pre-slip model,

like the cascade-up model, centers on processes that take place along major faults that already exist. An underlying assumption that gives this model importance is that monitoring increasing aseismic slip may be utilized as an indicator of an impending major earthquake. However, it has not yet been feasible to predict major earthquakes in natural settings by observing previous seismic slip.

The progressive localization framework is distinct in that it does not concentrate on processes that are restricted to pre-existing faults (Lyakhovsky et al, 1997). It instead explains the gradual progression from dispersed failures in a rock volume to concentrated deformation, culminating in the production of primary slip zones and major earthquakes. There are various clusters of seismicity throughout the localization process in a zone with many faults of varying scales. Each cluster may have its own foreshocks, one of which may initiate the mainshock rupture.

Foreshocks that occur near to mainshocks in space and time have been used to predict mainshocks because they are the most visible precursors. Foreshocks can be used as predictors that a particular source location for a main-shock is approaching stress, slide, or strength change. However, these foreshocks can only be defined retroactively by statistical studies of seismic catalogs that contain the mainshocks. In addition, not all large earthquakes are preceded by a foreshock series (Wu et al., 2014). Though some studies have show that certain big earthquakes along plate boundaries and in crustal fault systems were preceded by an elevated rate of sesimicity in the months to days preceding the mainshock (Bouchon et al., 2013; Tamaribuchi et al., 2018).

A component of each of these models to predict large earthquake is the clustering of earthquakes within a region both in time and in location. In the next section, we take a deeper look at earthquake clustering and the methods for quantifying and understanding clustering with the goal of improving models from both the localization approach discussed in this section and the previous section on statistical models.

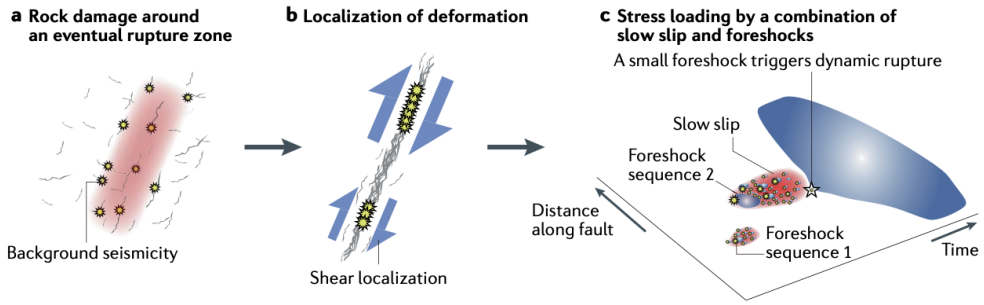

Figure 1: **Schematic illustrations of generation processes of large earthquakes. a** Progressive localization of shear deformation and background seismicity around a large rupture zone. **b** Shear localization and several foreshock sequences before the instability leading to the large rupture. **c** A space–time diagram of step-like increase in fault slip before a major earthquake associated with combined slow slip and foreshocks. A final rapid local loading by a small foreshock triggers the subsequent major dynamic rupture and circumvents the large nucleation process of a large patch. White and yellow stars denote epicentres of mainshocks and other events, respectively. As an example, two foreshock sequences accompanied with slow slip are displayed. (Kato & Ben-Zion, 2021)

2.4 Earthquake Clustering

A basic characteristic of seismicity is earthquake clustering. Clustering is observed most clearly as the concentration and frequency of earthquakes around large faults and tectonic plate boundaries and after large earthquakes (Scholz, 2002; Utsu, 2002). To explore clustering partition seismicity into groups closer in space and time than expected in a purely random distribution and look for patterns (Zaliapin & Ben-Zion, 2016). Understanding seismic clustering plays an important role in many elements of seismicity, such as identifying the characteristics and interactions of active fault structures(Zaliapin & Ben-Zion 2013a; Vidale & Shearer 2006; Vidale et al. 2006; Zhang & Shearer 2016).

2.4.1 Classification and Characteristics of Earthquake Clusters

Earthquake clustering is an important aspect of seismic activity that reveals valuable information about earthquake behavior. Clustering occurs in space, time, and size (such as magnitude and energy), and is the primary form of seismic activity. Spatial clustering appears as concentrations of earthquakes along major tectonic plate boundaries

and regional fault networks, while temporal clustering involves increased seismic activity following large earthquakes, leading to aftershock sequences (Omori,1894; Utsu, 1961; Utsu et al., 1995; Kisslinger, 1996). However, there is no formal definition of seismic clusters, which hinders systematic global analysis. This section reviews the development of an objective and reliable method for analyzing seismic clusters (Zalipin & Ben-Zion, 2013).

Mathematically, earthquake clustering is the separation of seismicity into groups that are closer in space and time than would be expected in a distribution that is completely random. These clustered groups of earthquakes reflect various triggering mechanisms, swarms and other forms of clustering as well as the traditional aftershock series.

A preliminary investigation of seismic clustering in southern California, detailed in the following section, revealed that clustering is strongly influenced by the crust's physical characteristics, leading to two primary types of clusters: "burst-like" and "swarm-like" (Zaliapin & Ben-Zion, 2013b). These findings demonstrate the importance of region-specific factors in understanding earthquake dynamics and improving seismic hazard assessments. While high-quality data for southern California made these trends apparent, global data suffers from lower quality catalogs with increasing magnitudes of completeness/reporting and location uncertainty, affecting cluster characterization.

In "A global classification and characterization of earthquake clusters", Zaliapin and Ben-Zion (2016) develop statistical tools for working with low-quality catalogs that are not sensitive to catalog inaccuracies. Zaliapin and Ben-Zion reveal that a main driving factor behind the types of clustering seen in a region is the effective viscosity of the region. Regions with increased effective viscosity experience burst-like clusters and regions with decreased effective viscosity experience swarm-like clusters. They show that local heat flow is the primary factor controlling the significant spatial dependencies of global earthquake clustering. Burst-like clusters are typical of cold places (mostly shallow seismicity of subduction zones), whereas swarm-like clustering is typical of hot regions (mainly mid-oceanic ridges). It is also shown that the sort of

plate-boundary deformation only has a little impact on the seismicity cluster style. The global findings summarized below are consistent with previous regional findings based on higher-quality data in southern California and these results are found to be resilient to the known catalog uncertainties and inadequacies.

The authors used a global earthquake catalogue produced by the Northern California Earthquake Data Center (NCEDC) for the period of 1975 to 2015, containing 256,993 events. They used a minimum magnitude of 4, which is higher than the completeness magnitude in some regions but their analysis technique is still robust.

Heat flow data from a study by Bird et al. (2008) was used to map the heat flow within seismically active areas and over the entire Earth surface. The highest heat flow production, reaching $0.3Wm^{-2}$, is found along the oceanic spreading ridges.

The global strain rate field data from Kreemer et al. (2014) was used to study the deformation of the Earth's surface. The analysis is based on the second invariant of the strain rate tensor (I_2) and the tensor style (S) defined by Kreemer et al. The tensor style is used to classify the type of displacement into contraction ($S < -0.5$), strike-slip ($0.5 < S < -0.5$), and extension ($S > 0.5$).

The minimum magnitude cutoff, m_c, in a catalog can affect aftershock analysis as earthquakes below this cutoff cannot be considered aftershocks of larger ones. A Δ-analysis, which only considers mainshocks with a magnitude of at least $m_c + \Delta$ (in this case, 4) and aftershocks 2 units below the mainshock magnitude, is used to compare aftershocks of mainshocks with different magnitudes. Aftershocks found through this method are called Δ-aftershocks and differ from regular analysis that considers all events.

To understand earthquake clustering first consider generalized earthquake distance. When working with a catalogue where each event i is characterized by its occurrence time t_i, hypocentre (ϕ_l, λ_i, d_i), and magnitude m_i, Baiesi $\&$ Paczuski (2004) define the proximity η_{ij} of earthquake j to earthquake i as:

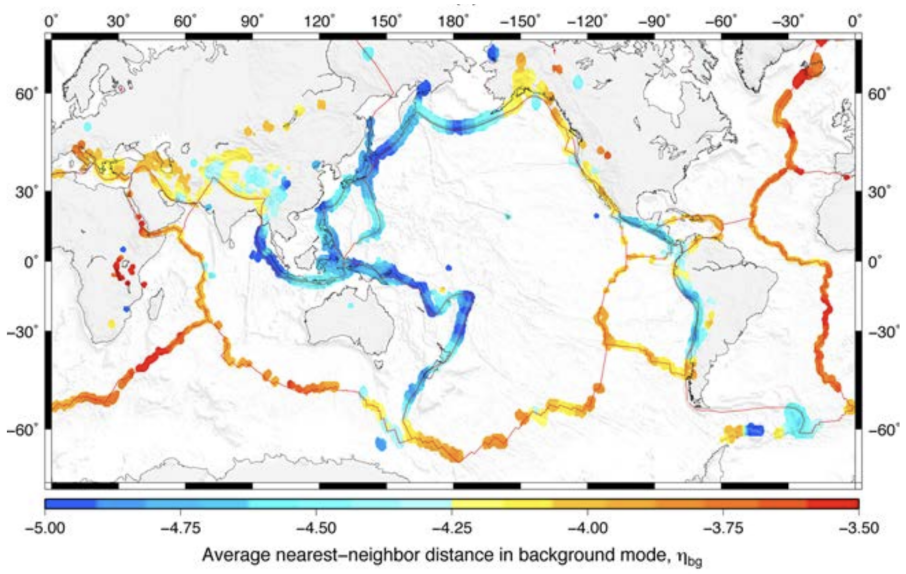

Figure 2: Global maps of selected parameters of seismic clustering. Average nearest-neighbour distance $log_{10}\eta_{bg}$ in the background mode. (Zaliapin and Ben-Zion, 2016)

$$\eta_{ij} = \begin{cases} t_{ij}(r_{ij})^d 10^{-bm_i}, & t_{ij} > 0. \\ \infty, & t_{ij} \leq 0. \end{cases} \tag{15}$$

The event intercurrence time, $t_{ij} = t_j - t_i$, is positive if event i occurs before event j; the spatial distance between earthquake hypo-centers is $r_{ij} \geq 0$, the (possibly fractal) dimension of the hypo-centers or epicenters is d, and b is the parameter of the Gutenberg–Richter law. Zaliapin & Ben-Zion (2013a, 2015, 2016a) address this proximity measure's justification and characteristics.

In identifying and characterizing earthquake clusters, it is necessary to identify the parent and offspring earthquakes in a catalog. To determine the distinct nearest neighbor(parent) j for each event i use the distance determined by (16), and indicate this distance by the same symbol, η_{ij}. For this definition, event i is referred to as an offspring of j, each event has a unique parent, aside from the first one in the catalog, and each parent might have multiple offspring.

Zaliapin & Ben-Zion (2016a) have shown that the distance to the nearest-neighbor is distributed bimodally. The distance in space and time normalized by magnitude of the

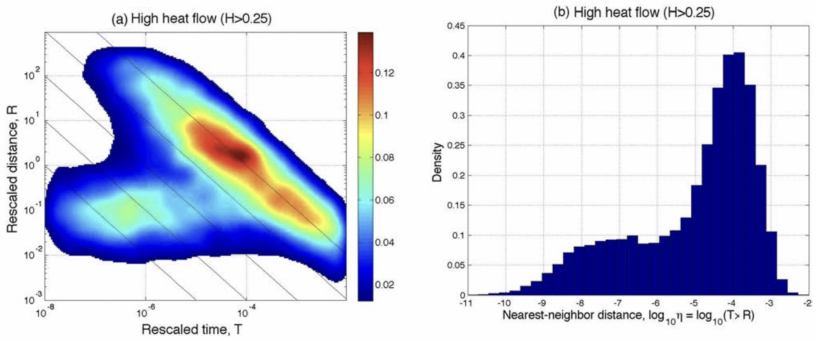

Figure 3: Generalized earthquake distance η and its normalized space and time components (T, R) in regions with high heat flow $H > 0.25$. (a) Joint distribution of the rescaled components (T, R) of the earthquake nearest-neighbour distance. (b) Distribution of the values of the nearest-neighbour distance η. Black diagonal lines in panel (a) depict levels of constant distance η (from top to bottom): $-\log_{10}\eta = 4, 5, 6, 7, 8$ (Zaliapin & Ben-Zion, 2016).

parent between a parent event j and its offspring event i is given by the formula

$$T_{ij} = t_{ij}10^{-qbm_i}; R_{ij} = (r_{ij})^d 10^{-pbm_i}; q + p = 1 \tag{16}$$

as defined by Zaliapin et al. (2008). Theoretically, it can be shown that events with Gutenberg-Richter magnitudes and time-stationary space-inhomogeneous Poisson flow have a uni-modal distribution of $(\log T, \log R)$ concentrated along a line $\log_{10} T + \log_{10} R =$ constant. In reality, observed seismicity behaves differently. It has been documented that detected seismicity over various regions has a bimodal joint distribution of $(\log_{10} T, \log_{10} R)$ where one of the modes coincides with background events and is comparable to the mode of a Poisson process (Zaliapin et al. 2008; Zaliapin & Ben-Zion 2011, 2013a,b, 2015, 2016; Gu et al. 2013; Davidsen et al. 2015; Reverso et al. 2015; Schoenball et al. 2015). The other mode is made up of clustered events located noticeably closer in space and time to their parents than is likely in a Poisson process.

A Gaussian mixture model approach can be used to distinguish the background and cluster modes in the bimodal distribution of earthquake distances. A Gaussian mixture model with two modes is used to intentionally select a nearest neighbor threshold η_0, so each event can be group into either the background (if $\eta_{ij} > \eta_0$) or cluster (if $\eta_{ij} < \eta_0$)

mode.

The distribution for a two mode Gaussian mixture model assuming that sample $x_i \in R^m, i = 1, ..., n$ is

$$F(x) = wN(x; \mu_1, \Sigma_1) + (1 - w)N(x; \mu_2, \Sigma_2). \quad (17)$$

Here, w is the mixture weight of the first mode and $N(x; \mu, \Sigma)$ denotes the Gaussian (Normal) distribution, with mean μ that is a vector with m elements and variance Σ that is a $m \times m$ positive-definite matrix. Model estimation, mode and background (cluster) events threshold assignment are discussed in detail in Zaliapin & Ben-Zion (2016a).

From this model, two important parameters of clusters, η_{bg} and Q, can be estimated for a region. The mean generalized earthquake distance η_{ij} of the background events from parent events is η_{bg}. The average value of the mode assignment probability $v = \max(w, 1 - w)$ over all events in a region is the regional mode separation quality Q. Q can take on any value between 0.5 and 1. If each event can be classified as part of one of the modes with probability 1 , $Q = 1$, while if the classification of each event to either mode has probability 0.5 then $Q = 0.5$.

Clustering identification is done using tree graphs and spanning forests. A spanning network representing a cluster is created by connecting each earthquake to its nearest neighbor, resulting in a tree graph (Zaliapin & Ben-Zion 2013a; Baiesi & Paczuski 2004). Applying a threshold $\eta \geq \eta_0$ removes links representing large distances between parent and offspring, resulting in a spanning forest where each tree represents a separate cluster. Trees may contain only single events (*singles*) or multiple events (*families*).

Events in each family are classified into *foreshocks*, *main shock* and *aftershocks*. The main shock is identified as the earthquake with the largest magnitude in the family, or the first to occur if there are multiple earthquakes with the same largest magnitude. Foreshocks are the events preceding the main shock, and aftershocks are the events following the main shock. Single events are classified as main shocks.

Earthquake clustering is space-dependent and related to heat flow production, with

the highest seismic activity occurring in subduction zones and the average heat flow having the highest values in extension environments, particularly along mid-oceanic spreading ridges. The behavior of earthquake clustering is driven by space-dependent heat flow production. Notable differences in earthquake clustering between high and low heat flow regions include higher earthquake intensity in low heat flow regions with predominantly contraction and transform deformation styles, lower proportion of background events in low heat flow regions, and faster time decay of cluster events in high heat flow regions leading to stronger time separation between background and cluster modes.

A 1-D Gaussian mixture model applied to the nearest-neighbor distances of events within circles centered at the epicenters of all examined earthquakes was used to estimate the space-dependent threshold η_0 that separates the cluster and background modes. This threshold was then used to partition events into cluster and background populations. From this analysis properties of the cluster characteristics η_{bg} and Q were illuminated. The highest earthquake intensity and lowest values of η_{bg} were observed within convergent environments, which flexibility large spatial variability and intermittence in separation quality Q. The lowest earthquake intensity and largest values of η_{bg} were observed along divergent boundaries, and these boundaries also have high mode separation quality Q. Intermediate values of earthquake intensity and background position η_{bg} were observed along transform boundaries, along with the mode separation quality Q showing high intermittency. A GLM approach and Spearman's correlation were used to compare earthquake cluster statistics with strain rate tensor and heat flow parameters. This approach further documented the correlation between heat flow and these examined cluster characteristics (Zalipain & Ben-Zion, 2016).

A key property of earthquakes clusters uncovered in this analysis was that cold areas have much larger clusters than hot areas. The study examines a catalogue of 256,993 events and partitions them into 135,840 clusters using a specific procedure. Of these clusters, 85.6$\%$ are single events and 14.4$\%$ are families with sizes ranging from 2 to

6584. The distribution of cluster size is found to be different in areas with high and low heat flow levels, with a power law tail approximation of

$$S(N) = \mathsf{Prob}[\text{cluster size} > \mathsf{N}] \propto \mathsf{N}^{-\alpha} \tag{18}$$

with $\alpha \approx 2$ in hot areas and $\alpha \approx 1$ in cold areas. The observed difference in cluster size distributions implies that cold areas have much larger clusters and a larger proportion of clusters with size $N > 10$. The dominance of large clusters in cold regions is explained by statistically higher maximal magnitude and better quality of catalogs in cold regions compared to hot ones. The cluster size is stochastically larger in cold regions for clusters with main shock magnitude $m > 6$, while the size of intermediate-magnitude clusters (with main shock magnitude $m < 6$) is stochastically larger in hot regions. In addition, the proportion of smallest clusters (singles) among all detected clusters is higher in cold areas. The probability of being a single is higher for small-magnitude events. With the existing data, it is difficult to conclude whether this effect is related to the inferior catalog quality in hot regions or is a real physical property.

The proportion of foreshocks among foreshocks and aftershocks, p_F, is higher in areas with high heat flow, typically with values greater than 0.2, while in areas with low heat flow the typical proportion is very small, $p_F < 0.1$. This increased production of foreshocks in hot regions is supported by the analysis of the value of p_F averaged for different combinations of strain rate tensor's style S and second invariant I_2. Spatial patterns similar to those reported in 4 are also seen for other examined cluster characteristics.

When looking at the worldwide distribution of the aftershock magnitude gap ΔA defined for families with aftershocks as the difference between the magnitudes of the main shock and the largest aftershock, this gap is generally larger within cold regions, with typical value of $\Delta A \approx 0.8$, while in hot regions it is typically smaller, $\Delta A \approx 0.55$. The comparison of p_F and ΔA with heat flow and strain rate tensor parameters using Spearman's correlation and GLM approach, it is notable that the values of the aftershock

magnitude gap reported here are lower than the value $\Delta A \approx 1$ suggested by the Bath law. This deflation is artificial and is due to the fact that we consider families with main shock magnitude $m \geq 5$, which is only one unit above the magnitude cut-off $m_{min} = 4$ selected for this study. The magnitude gap is affected by the catalog completeness magnitude, since a higher completeness magnitude leads to smaller observed values of ΔA. The reported difference in magnitude gap might be influenced to some extent by inferior catalog quality in hot areas. However, Zaliapin & Ben-Zion(2013a) reported lower magnitude gap in hot regions in a local study in southern California, where the quality of catalogs is comparable in both cold and hot regions. The study believes that the magnitude gap difference between hot and cold areas is a real phenomenon that will be confirmed in future studies with better catalog quality.

As mentioned previously, the structure of earthquake families is represented by a tree, T, with vertices, $V = v_i, i = 1, ..., N$ representing earthquakes and edges, e_i connecting them to their parent within the same family. The first event in the family is referred to as the root and all other events have a single parent within the same family. Two statistics are used to study the structure of earthquake families: the average family branching and the average leaf depth. The average family branching is the average number of offspring per parental vertex of the tree, while the average leaf depth is the average number of edges between a leaf and the tree root. It is expected that the leaf depth and family branching are negatively correlated, meaning that as the leaf depth increases, the family branching decreases. Zaliapin and Ben-Zion (2013b) found that these statistics are strongly coupled with the heat flow in southern California, with leaf depth increasing and family branching decreasing as heat flow increases. The same trend is observed on a global scale. The values of these statistics also depend on the family size, which can affect spatial analysis. A least-square regression analysis suggests that the examined statistics have a relation to the family size in the intermediate size range of 5 to 20. The spatial distribution of these statistics also shows that cold regions typically have a smaller average leaf depth and a larger average family branch-

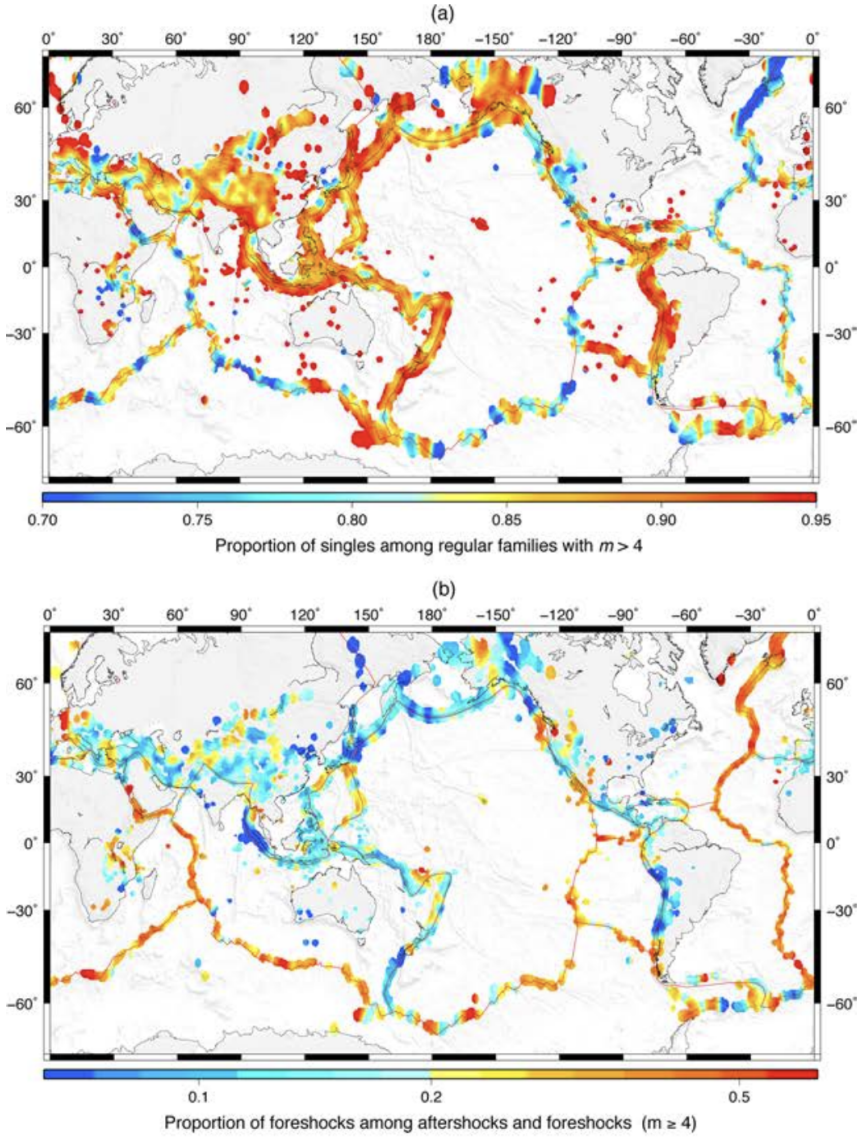

Figure 4: Global spatial distribution of selected earthquake cluster statistics. (a) Proportion p_S of singles among regular clusters. (b) Proportion p_F of foreshocks among foreshocks and aftershocks (Zaliapin and Ben-Zion, 2016).

ing than hot regions. The comparison of these statistics with heat flow and strain rate tensor parameters using Spearman's correlation and GLM approach confirms that heat flow exerts the primary control on the values of these two statistics.

This study investigates earthquake dynamics on a global scale and found that earthquake statistics and cluster characteristics have a spatially dependent distribution, which correlates with global heat flow production and style of lithospheric deformation indicated by an estimated strain rate tensor. The results support earlier findings that there is non-universal, region-specific behavior of seismicity, and suggest that there are two primary types of earthquake clustering: burst-like clusters characterized by a large main shock in cold regions, and diffuse clusters with a smaller main shock and a larger number of offspring events in warm regions. The findings are consistent with theoretical expectations based on a viscoelastic damage rheology model.

2.5 Applications - Earthquake clusters in southern California

2.5.1 Identification and stability

This section explores the application of the techniques discussed in the previous section to identify and characterize the clustering of seismicity in southern California. The study "Earthquake clusters in southern California I: Identification and stability" by Zaliapin and Ben-Zion (2013a) uses the method based on the bimodal distribution of nearest-neighbor earthquake distances in a combined space-time-magnitude domain, which allows for partitioning an earthquake catalog into separate individual clusters mentioned in the previous section. The clusters are divided into singles, which contain just one event, and families, which contain multiple events, and are subclassified into foreshocks, mainshocks, and aftershocks. Recall this method is characterized by its ability to use only three easily estimated parameters, its ability to uniformly analyze clusters associated with mainshocks of greatly different magnitude, its high stability with respect to parameters, minimal reported magnitude, catalog incompleteness, and location errors, and its absence of underlying assumptions or governing models for the

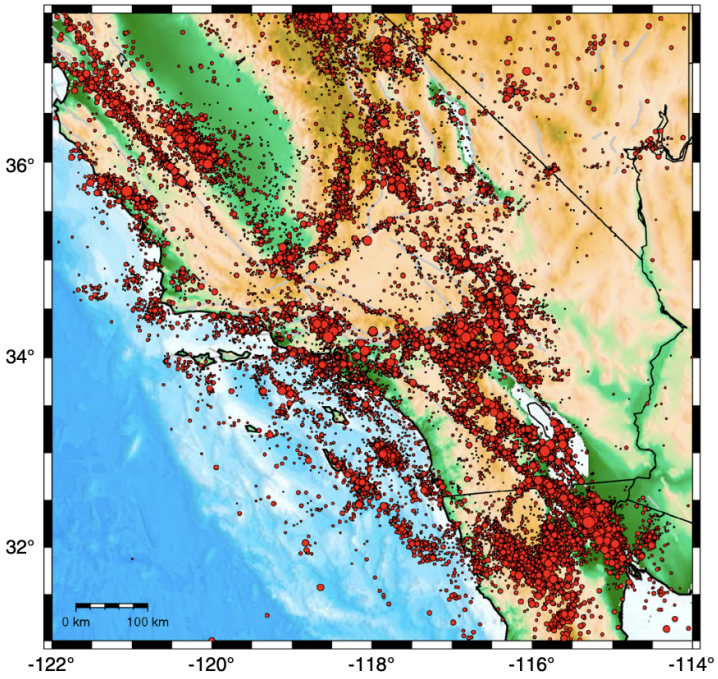

Figure 5: Map of earthquake epicenters, $m \geq 2$, from the relocated catalog of Hauksson et al. (2012). Circle size is proportional to magnitude. Major faults are shown by gray lines.

expected earthquake cluster structure.

The study uses a the by Hauksson et al. (2012) catalog. Figure 5 in the study shows the location of these earthquakes and Figure 6 provides a visual representation of changes in seismic intensity, which are mostly related to aftershocks of large earthquakes.

Despite using a lower completeness magnitude of $m_c = 2$ compared to the estimated

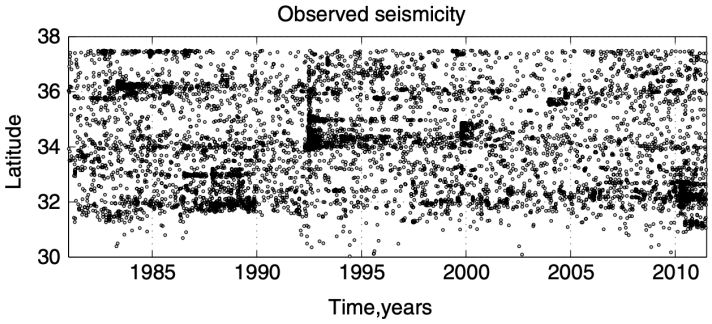

Figure 6: Epicenters of the earthquakes with $m \geq 3$ as a function of time and latitude.

completeness magnitude of 3.0 for southern California, the study found that the cluster structure of events is unaffected by incompleteness or the minimum reported magnitude (Felzer, 2008; Schorlemmer and Woessner, 2008). This indicates that the recovered cluster structure is similar to what would be observed in a complete catalog. Results for a higher magnitude threshold of 3 are similar, but there are insufficient clusters to obtain clear results.

In this study, earthquake clusters are identified by the nearest neighbor approach discussed in detail in the previous section. The analysis shows two distinct groups: clustered events that are close to each other in time and space, and background events that are farther apart than expected. Although background events are not entirely homogeneous, deviations are smaller than for clustered events. The study focuses on significant clusters, their properties, and the single largest event in each cluster, regardless of homogeneity. This approach differs from catalog declustering, which removes events to create a homogeneous set (Gardner & Knopoff, 1974). Recent research shows that the assumption of catalog homogeneity is often incorrect due to seismic migration and changes in activity (Luen & Stark, 2012).

Results from analyzing 2,105 events with $m \geq 3$ in southern California, using the Hauksson et al. relocated catalog, suggest that they form a Stationary Homogeneous Poisson (SHP) process. This process is a marked Poisson point process that is uniform across multiple dimensions, stationary over time, and has magnitudes that follow the Gutenberg-Richter distribution. The joint 2-D distribution of time and distance in SHP is unimodal and concentrated along the line $\log_{10} T + \log_{10} R = -5$, as shown by exploring Nearest Neighbor distances (NND). A histogram of NND on a logarithmic scale also displays a clear unimodal shape in the joint distribution of time and distance of the SHP process, which closely matches the theoretical Weibull distribution (Zaliapin et al., 2008; Hicks, 2011). This aligns with the analysis outlined in the previous section.

Observed seismicity in Southern California uncovers a prominent bimodal distribution of η and of the joint distribution of (T, R). The first mode, referred to as the

"background," is similar to the distribution of a SHP process. The second mode, referred to as "clustered," is located closer to the origin. This bimodal distribution has also been observed in global seismicity and regional seismicity in other areas such as Nevada, Japan, New Zealand, and Africa (Hicks, 2011; Bautista, 2011). The bimodal distribution is caused by groups of earthquakes happening within localized regions in space and time, mainly corresponding to foreshock-mainshock-aftershock sequences or swarms. The bimodality of the (T, R) distribution is used to identify individual space-time clusters of seismicity.

A nearest-neighbor spanning network from the southern California earthquake catalog was created by connect each earthquake to its closest neighbor to form a single cluster that includes all earthquakes. This nearest-neighbor network is a tree structure without loops (Baiesi & Paczuski, 2004). The links in the tree are then separated into strong and weak links based on their proximity, NND η. Weak links correspond to the background part of the bimodal distribution of seismicity, while strong links correspond to the clustered part of the bimodal distribution. The weak links are removed to form a spanning forest, which is made up of a collection of distinct trees, which includes single-event trees (singles) and multievent clusters (families) as discussed in the previous section. 60% of the earthquakes have strong links and form families, while 40% have weak links and are singles in the southern California catalog (Zaliapin & Ben-Zion, 20213a). If we zoom in on a specific aftershock sequence we can see the highly clustered nature of families of events, while singles are more uniformly distributed.

When examining the internal composition of the clusters within the nearest-neighbor spanning forest, families, are separated into "mainshocks," "aftershocks," and "foreshocks." The type of earthquake depends on the time and magnitude. The analysis shows that the proportion of different types of earthquakes is stable for families with magnitudes below 5, with about 37% of mainshocks and singles, 56% aftershocks, and 7% foreshocks. For larger earthquakes, there is a preference for mainshocks. The spatial and temporal distribution of mainshocks and singles is larger than that of the clusters

identified by the nearest-neighbor analysis. The mainshock/single field is not Poissonian and the cluster identification procedure does not distort the nonhomogeneous and possibly nonstationary background events. However, the clustering of mainshocks and singles lacks bimodality, so additional rules are required to analyze such data (Zaliapin & Ben-Zion, 2013a).

The technique for detecting earthquake clusters used in this study is based on the earthquake distance equation (Equation 16) and a cluster threshold η_0 as described in the previous section. The technique is completely parameterized by three values $(b, d_f, \text{and} \eta_0)$ that are estimated from observations(Marzocchi & Sandri, 2003; Harte, 1998; Kagan, 2007; Molchan & Kronrod, 2009; Hicks, 2011). The performance and stability of the technique were tested using catalogs generated by the ETAS model(Ogata, 1998). The results suggest that the technique can accurately identify clusters in the ETAS model, is stable with respect to various parameters, and can be applied to the observed catalog of southern California seismicity. The study uses a version of the ETAS model that assumes isotropic spatial kernel and homogeneous spatial background, which is commonly used in seismicity analysis(Veen and Schoenberg, 2008; Wang et al., 2010; Chu et al., 2011). The results suggest that the cluster structure in the southern California catalog is similar to that generated by the ETAS model, and the proposed technique can robustly recover the cluster structure.

When analyzing the statistics of the detected earthquake clusters, Zaliapin and Ben-Zion had two objectives. The first objective was to validate the proposed cluster detection technique by reproducing known statistical features of aftershocks and foreshocks. The second objective was to uncover new properties of the earthquake clusters.

When looking at the magnitude distributions of mainshocks/singles, aftershocks, and foreshocks, the cumulative proportion of earthquakes above or equal to magnitude m is normalized by the magnitude, $[1 - F(m)]x10^m$ where $F(m)$ represents the empirical cumulative distribution function of magnitudes, to emphasize the changes in the exponential index. This transformation shows that all three distributions are exponen-

tial within the magnitude range of 2.5 to 4.5, with a lower index, $b \approx 1$, for mainshocks and aftershocks and a higher index, $b > 1$ for foreshocks. The estimation of the b-value, which is a measure of the rate of earthquakes, confirms the visual observations from the cumulative plots of observed seismicity (Tinti and Mulargia, 1987). The b-value for mainshocks and aftershocks is the same, while it is higher for foreshocks. The mainshock/single magnitude distribution does show an upward deviation from the exponential distribution and it is thought this is due to growth of stress concentration in elastic solid with rupture size, but the exact reason remains unclear, and may be due to either statistical artifacts or physical processes. The results suggest that the behavior of natural seismicity clusters and foreshocks, in particular, cannot be fully explained by the ETAS framework (Zaliapin & Ben-Zion, 2013a).

When the number of offspring events (N_{off}) in a catalog of observed earthquakes is compared to an ETAS model catalog, the results showed that the N_{off} distribution for fixed m in the observed catalog deviates significantly from a Poisson distribution and is better approximated by a negative binomial distribution, which is consistent with previous research (Kagan, 2010).

For observed seismicity in southern California, the offspring number scales with the event magnitude as

$$N_{off} \propto 10^{cm}, c = 0.93 \pm 0.06 \tag{19}$$

where the estimation is done within the range $4 \leq m \leq 6$, and at the 95% confidence level.

The estimated mean and variance of N_{off} for the observed catalog is larger compared to the ETAS model. In particular, the average N_{off} for earthquakes with magnitude less than 4 is significantly smaller in the ETAS model than in the observed catalog. The ratio between the variance and average also increases from 1 to about 100 as the magnitude of earthquakes increases from 2 to 6 in both the observed catalog and ETAS model, further emphasizing the inadequacy of the Poisson model for describing the

N_{off} distribution (Zaliapin & Ben-Zion, 2013a).

An important result of this increased variability of the N_{off} distribution in the observed catalog is the existence of a large population of "singles" or main shocks with no offspring. This observation cannot be solely explained by catalog artifacts such as incompleteness or the minimal reported magnitudes. Singles make up 84% of all detected clusters, 53% of clusters with magnitude greater than or equal to 3, and 17% of clusters with magnitude greater than or equal to 4. The largest single event detected has a magnitude of 5.0 (Zaliapin & Ben-Zion, 2013a).

The distribution of cluster size N is found to be closely approximated by a Pareto distribution with a index $a \approx -1$, which can be represented by the combination of exponential mainshock magnitude distribution and the exponential number of offspring for a given mainshock with index $\alpha \approx b$ (Saichev et al., 2005). The number of aftershocks per cluster is larger and scales with the mainshock magnitude as $N_A \propto 10^{\beta m}$ with an index of $\beta \approx 0.99 \pm 0.06(95\%CI)$. The number of foreshocks doesn't exhibit a clear exponential scaling but the best exponential fit would have a smaller index of $\beta \approx 0.6$. The increase of fore/aftershock number with the cluster mainshock magnitude is due to the existence of the catalog lower cutoff magnitude. The cluster size is found to be independent of the mainshock magnitude and robust with respect to the earthquake magnitudes after conducting various tests (Freedman, 2005; Kruskal & Wallis, 1952).

To understand the temporal structure of families the intensity of earthquakes in the clusters versus time is averaged over all the detected clusters in both regular and Δ-analyses, specifically looking at the relationship between the intensity and time relative to the main shock. Results from both regular and Δ-analyses (which only includes aftershocks and foreshocks within Δ magnitude units from the main shock) showed a conventional pattern of foreshock-mainshock-aftershock, with lower number of foreshocks and higher number of aftershocks. The intensity of both aftershocks and foreshocks decreases over time, with a power-law decay seen in both. Results were consistent with the Omori-Utsu law (13) for the intensity of both foreshocks and aftershocks, and

showed that the productivity index K is a constant dependent on Δ but not the main shock magnitude in Δ-analysis, while in regular analysis it scales with the main shock magnitude m as

$$K = 10^{\beta m} \tag{20}$$

with $\beta \approx 1$ for aftershocks and $\beta < 1$ for foreshocks. The difference between the aftershock intensity in regular vs Δ-analysis was significant, while the difference between the foreshock intensities was much smaller. This difference was seen in the distribution of magnitude differences between the main shock and the family events, with foreshocks having a closer magnitude to the main shock compared to aftershocks. However, the distributions of the magnitude difference between the main shock and the largest foreshock or aftershock were statistically the same. These results suggest the existence of an accelerated failure process as the time of main shock approaches (Mogi, 1969; Keilis-Borok & Kossobokov, 1990; Bufe and Varnes 1993; Ben-Zion & Lyakhovsky,v2002; Turcotte et al., 2003).

Previous research has shown a systematic difference between the magnitudes of mainshocks and their largest aftershocks with an average magnitude difference close to 1.2 (Båth, 1965; Kisslinger & Jones, 1991; Shcherbakov & Turcotte, 2004; Shearer, 2012). The current analysis of southern California confirms this, with an average magnitude difference of 1.1 for aftershocks and 1.2 for foreshocks. The differences in magnitude, $\Delta_m = m_{\text{mainshock}} - m_{\text{largest-event}}$ are found to have an almost uniform distribution within the range of $[0, 2]$. However, the data shows significant deviations from a uniform distribution in the foreshock magnitude distribution in the ETAS model. Further testing is needed to determine if this deviation is systematic and if the ETAS model deviates from observations regarding foreshock magnitudes.

The size of earthquake clusters relates to the main earthquake magnitude, with the area A defined as the smallest convex hull containing them. Aftershock areas scale with the mainshock magnitude m as $A \propto 10^{\gamma m}$, $\gamma \approx 1$, and are independent of family size

N, suggesting a damage zone around the main rupture that scales with its magnitude. Foreshock areas are an order of magnitude smaller on average than aftershock areas and independent of family size. Cluster duration is independent of main earthquake magnitude and slightly increases with family size, indicating elastic stress transfer dominates stress relaxation after a mainshock.

Seismic clustering in southern California is tied to crustal properties and changes over tens of kilometers. Two primary types of clusters exist: ”burst-like” and ”swarm-like.” Burst-like clusters have a distinct large main shock, few foreshocks, and more first-generation offspring, indicating brittle collapse in regions of low fluid and heat flow. The Mojave, Ventura, and San Gabriel regions have burst-like clusters. Swarm-like clusters lack a noticeable primary shock, have more foreshock activity, and many secondary offspring, indicating mixed brittle-ductile failure in regions of high fluid, heat, and/or soft sediment flow. The Salton Sea and Coso geothermal districts have swarm-like clusters. Effective viscosity is the primary driver, with increased viscosity leading to burst-like clusters and decreased viscosity leading to swarm-like clusters. These trends in southern California provide valuable insights into earthquake dynamics and seismic hazard assessments but are limited by the quality of global earthquake data, which can be improved using statistical tools developed by Zaliapin and Ben-Zion (2016) discussed in the previous section.

2.5.2 Spatial Variations of Rock Damage Production

Earthquake ruptures result in rock damage within their source volumes, leading to an increase in crack density and changes in various rock properties such as elastic moduli, mass density, seismic velocities, attenuation, anisotropy, permeability, and conductivity (Lockner et al., 1977; Dresen & Gueguen, 2004; Mavko et al., 2009). The ways in which rock damage progresses reflects important underlying dynamics that precede large brittle instabilities and an accurate model of seismicity would take these elements into account (Peng & Johnson, 1972; Hamiel et al., 2004; Renard et al., 2018).

”Spatial variations of rock damage production by earthquakes in southern California” (Ben-Zion & Zaliapin, 2018) endevours to estimate the amount of rock damage produced by earthquakes in Southern California. The authors use data from earthquakes with magnitudes in the range $[2, 4)$ from 1981 to 2017 (Hauksson et al., 2012). The study aims to understand the relative production of rock damage in different parts of Southern California. The authors use theoretical relations from earthquake phenomenology and fracture mechanics to estimate the production of fracture area and volume generated by observed seismicity. The results of the study indicate that there is a zone with ongoing damage production between the Imperial fault and the Eastern California Shear Zone. The regions around the 1992 Joshua Tree, Landers and Big Bear earthquakes are active before 1990 and outline the future earthquakes. The seismicity and damage zone become more pronounced and continuous with increasing depth leading up to these larger events. Finally, implications of these results for the properties and dynamics of the plate-boundary region in Southern California are discussed.

Estimating the amount of rock damage produced by earthquakes in southern California is done by using observed seismicity, basic relations from earthquake events and fracture mechanics (Jamtveit et al., 2018). The authors provide theoretical formulations for calculating the fracture area and rupture volume, which are then implemented using the earthquake catalog for the period 1981-2017.

The total number of earthquakes with a magnitude of M or greater is given by the Gutenberg-Richter exponential relation (Gutenberg B. & Richter C. 1954). The fracture area of each earthquake is estimated by assuming it can be approximated as a circular crack with a uniform strain drop in a solid (Eshelby, 1957; Ben-Zion, 2008; Ross et al., 2016). The seismic potency of each earthquake is related to its magnitude (Hanks and Kanamori, 1979). By combining these relationships, the total fracture area is calculated by integrating the fracture area of each earthquake. The calculation shows that the smallest earthquakes in the population dominate the total fracture area.

A method for estimating the rupture volume generated by earthquakes in the magni-

tude range $M_1 \leq M \leq M_2$ by considering the width (thickness) of the rupture zone for each earthquake is outlined in this paper. The width of the rupture zone is proportional to the rupture radius and is related to the dynamic stress intensity factor and the ratio of stress drop to strength drop (Ben-Zion & Ampuero, 2009). Using this information, the damage volume generated by each individual earthquake can be calculated. Integrating the results over the magnitude range gives the total rupture volume, which is dominated by the largest earthquakes included in the analysis.

The above techniques were implemented to estimate the relative production of fracture area and rupture volume in different parts of Southern California caused by earthquakes associated with ongoing background activity. This analysis uses a catalog of earthquakes from 1981 to 2017, and only earthquakes with magnitude $2 \leq M < 4$ to obtain results that are representative of a typical inter-seismic period for all faults in the study area. They use a grid of 300×300 with an average spacing of 10 km to calculate the cumulative damage volume caused by the earthquakes, then divide it by the time interval duration of 37 years to obtain damage values in units of km^3 per year. Finally, they smooth the resulting map using a Gaussian filter with a standard deviation of 0.15 degrees (equivalent to 16 km).

To assess the temporal stability of the estimated damage production, compare the damage volume production before and after 1990, the year in which the largest event occurred in the study area, which was the 1992 M7.3 Landers earthquake. The comparison is done by calculating the proportional change of damage volume

$$\Delta_{\text{volume}} = \frac{V_{\text{after}} - V_{\text{before}}}{\max(V_{\text{after}}, V_{\text{before}})} \tag{21}$$

Where V_{before} represents the damage volume production rate prior to 1990 and V_{after} represents the value for the same rate after 1990. The change in damage volume production rate (Δ_{volume}) after 1990 has a range of -1 to 1, with negative values indicating a decrease and positive values indicating an increase.

Southern California has concentrated background damage in several areas, includ-

ing the San Jacinto Fault Zone, Brawley Seismic Zone, South Central Transverse Ranges, Eastern California Shear Zone, and Elsinore Fault. The rock damage is not uniformly distributed along the fault structures but instead focused in persistent active areas, which are major damage hotspots. Most moderate to large earthquakes occur within these hotspots, which are also connected to low magnitude seismic events. However, there are also active damage zones that have not experienced any moderate to large earthquakes in the last 30 years.

Rock damage production has remained concentrated in the same zones before and after 1990, with some fluctuations indicating a slow migration of active patches and changes in damage intensity. Isolated patches have shown decreased damage production after 1990. Overall, background events' damage production is stable across the examined region, despite the earthquake process's complexity and the shortness of the catalog (Ben-Zion & Zaliapin, 2019).

Fracture area maps were created using the same approach as the damage volume results, and they are nearly identical, confirming that the analysis represents inter-seismic activity and is not affected significantly by sample artifacts. The numerical ratio of estimated total rupture area and volume changes with the lower and upper boundaries of magnitude. It increases when the lower boundary extends to smaller events and decreases when the upper boundary extends to larger events (Ben-Zion & Zaliapin, 2019).

Results from damage volume production analysis suggest stable damage production by background events in space and time, with active regions around the 1992 M6.1 Joshua Tree, M7.3 Landers, and M6.3 Big Bear earthquakes before 1990 outlining the ruptures of future events(Ben-Zion & Zaliapin, 2019).

Changes in rock damage in southern California were explored by applying basic theoretical principles for the area and volume of rupture caused by earthquakes, which follow the Gutenberg-Richter frequency-magnitude relation. Results show ongoing occurrences of background earthquakes in certain regions, with the largest continuous region associated with the San Jacinto fault zone. Rock damage maps suggest a possi-

ble large-scale active seismic zone connecting the Imperial fault and Brawley seismic zone in the south with the Elsinore fault zone to the north. Rupture zones of past earthquakes were found to have ongoing background seismicity before the events occurred. These results provide valuable insights into the seismically active configuration of the plate boundary in southern California (Ben-Zion & Zaliapin, 2019).

2.5.3 Artifacts of Earthquake Location Errors and Short-term Incompleteness

Inconsistencies in data collection can impact analysis towards the goal of quantify earthquake clustering. Quantitative characterization of spatio-temporal earthquake clustering and its relationship to the physical properties of the lithosphere, as well as its response to various natural and human-induced loads has been a fundamental goal of seismologists (Enescu et al. 2009; Holtkamp et al. 2011; Brodsky & Lajoie 2013; Ellsworth 2013; Gu et al. 2013). Advancements in seismology have enabled the study of smaller magnitude earthquakes using innovative statistical methods and improved catalog data (Hainzl 2013; Hainzl et al. 2014; Moradpour et al. 2014). This has opened up new avenues for exploration, such as the structure of seismic bursts and induced seismicity, which cannot be investigated using data from larger earthquakes alone (Vidale & Shearer 2006; Chen & Shearer 2013; Eaton et al. 2014). However, the increased focus on smaller earthquakes has also led to various catalog uncertainties that can impact result (Kuge 1992; Rohm et al. 1999; Storchak et al. 2000; Kagan 2003). ”Artefacts of Earthquake Location Errors and Short-term Incompleteness on Seismicity Clusters in southern California” (2015) by Zalipain and Ben-Zion focuses on the effects of earthquake catalog uncertainties, such as event location errors and short-term incompleteness, on the estimated earthquake clustering and triggering. The paper uses the method of estimating earthquake cluster properties outlined in section 2.4.1 and the relocated earthquake catalogs in southern California to explore these uncertainties. The results showed that the method can identify thousands of earthquake clusters, including small-to-medium magnitude events, in different seismic environments, and classify them into

three main types. However, extending these results to other seismically active areas is challenging because of the non-uniform quality of typical data. This study investigates the effects of catalog errors on inferred cluster properties and documents the striking seismic patterns that arise as a result of these errors.

This research uses three earthquake catalogs from southern California for analysis, Hauksson et al. (2013), Richards-Dinger and Shearer (2000), and ANSS catalog(2015).

The earthquake clustering method outlined in section 2.4.1 was used in this study to identify and analyze parent–offspring pairs of earthquakes from these three catalogs.

Geography strongly affects the location errors of earthquakes in the Southern California HYS catalog from 1981-2013. The errors decrease in central southern California compared to peripheral areas due to differences in the seismic network quality. Three statistics - number of P and S picks used, number of differential times, and number of similar events - can evaluate network quality. The number of picks used shows the strongest correlation with location error, but the distribution doesn't match that of the other two statistics, which have higher values in high seismic activity areas (Zaliapin & Ben-Zion, 2015).

Location errors in earthquakes significantly impact the spatial relationship between parent and offspring events. Analyzing the rescaled time and distance between events showed that high location errors increase the distance-to-parent, particularly when the parent event is large. Location errors can cause related artifacts, including decreased spatial decay rate of offspring and artificial increase in the spatial distribution of foreshocks. These effects are not limited to the study's specific parent identification method but also expected in other cluster approaches (Zaliapin & Ben-Zion, 2015).

The second artifact is underestimated offspring production. When analyzing the impact of mislocations in earthquake catalogs on the identification of parent-offspring relationships, the study defines close events as those separated by a combined distance of $\eta < 10^{-5}$. The proportion of events with close offspring and close parent are examined as a function of the absolute horizontal location error. The results show that as

the error increases, the proportion of events with close offspring and close parent decreases, leading to an underestimation of offspring productivity and the total number of clustered events. When comparing the location errors to the estimated rupture lengths of the events and their parents the location errors are comparable to the half-rupture-length of the estimated parent for many events, potentially leading to incorrect parent identification (Zaliapin & Ben-Zion, 2015).

Location errors can cause miss-identification of background events as clustered events, resulting in an overestimated background rate. The study found an increase in the proportion of background events with large location errors and a three-fold increase in single events as absolute location error increases. While relative error exacerbates the effect, the miss classification is negligible (Zaliapin & Ben-Zion, 2015).

Small events following large earthquakes may not always be recorded in seismic catalogs, causing a change in b-value over time and space (Utsu et al. 1995; Wiemer & Katsumata 1999; Narteau et al. 2002; Kagan 2004; Lolli & Gasperini 2006; Helmstetter et al. 2007; Peng et al. 2007). B-value measures earthquake frequency at different magnitudes. To investigate this change, earthquake magnitudes were compared for events occurring soon after a large earthquake $(t < 1\text{day})$ and those far away $(t > 10$ days$)$ in Southern California. Magnitudes were also compared for rescaled times (T) of events. This analysis found that the b-value decreases by over 0.1 near a parent earthquake due to short-term incompleteness. This effect is more noticeable when using rescaled times. The average magnitude of earthquakes also decreases after a large earthquake, indicating an increase in b-value, which affects small and large magnitude earthquakes. This short-term incompleteness may bias regional b-value studies, even for magnitudes above the completeness threshold.

Next, four alternative catalogs of earthquakes in Southern California are compared and analyzed. These include the Hauksson et al. (2013) catalog, the Richards-Dinger and Shearer (2000) catalog, and two ANSS subcatalogs. The goal is to show that the inaccuracies and incompleteness in earthquake locations, seen in the Hauksson et al.

catalog, are not unique to it and are present in the other catalogs as well. This analysis aims to use the relationships between location errors and clustering to determine the accuracy of event locations in each of the four catalogs, ranking them based on the overall accuracy.

The study compares earthquake catalogs in Southern California to determine if reported location errors and short-term incompleteness are unique to the Hauksson et al. (2013) catalog. Four catalogs are compared: Hauksson et al. (2013), Richards-Dinger and Shearer (2000), and two ANSS subcatalogs. The joint distribution of the rescaled time and space components of the earthquake nearest-neighbor distance (η) and the $\log_{10} \eta$ in the Richards-Dinger and Shearer (2000) catalog resembles that of the Hauksson et al. (2013) catalog. The three main artifact studied in the Hauksson et al. (2013) catalog were also reproduced in the Richards-Dinger and Shearer (2000) catalog. The results suggest that these artifacts are present in all catalogs studied and are not specific to a single catalog or relocation method.

The analysis is done for two complementary groups of events, with and without magnitude constraints. As the quality of the catalogs decreases, the location of the events shifts towards higher rescaled times and distances, the separation of the clustered and background modes decreases, and the proportion of background events increases while the proportion of clustered events decreases. Other cluster statistics remain consistent among the alternate catalogs. The study also examines the effects of short-term incompleteness in the catalogs and finds that the average magnitude decreases with time after the parent and stabilizes after 10 days. The results provide a rough assessment of the size of fluctuations that might be expected as a result of varying location quality (Zaliapin & Ben-Zion, 2015).

In summary, two types of well-known uncertainties in the catalogs: location errors and short-term incompleteness, effects the estimation of earthquake cluster statistics. These uncertainties can significantly bias the results of cluster analysis and affect the estimation of earthquake background rates, triggering productivity, and b-value.

These uncertainties affect the estimated structure of small-magnitude earthquake clusters, causing many events to be misidentified as background seismicity and leads to overestimated background rates and underestimated clustering. Short-term incompleteness impacts the estimation of the b-value, which can be confused with magnitude dependence. The study focuses on cluster statistics related to parent-offspring pairs and shows that while large errors in individual parent-offspring identification do not propagate to the global cluster statistics, they still have a significant effect on the estimation of individual earthquake properties (Zaliapin & Ben-Zion, 2015).

2.6 Declustering

Using declustered catalogs (catalogs that do not include foreshocks, aftershocks and other strong forms of clustering) can uncover more subtle features of earthquake attributes and patterns (Dieterich, 1994; Console et al., 2006; Felzer & Brodsky, 2006; Lengliné et al., 2012; Ross et al., 2017; Kato & Ben-Zion, 2021).

The short duration of earthquake records and strong clustering of seismicity can mask other properties and hinder our ability to understand long-term earthquake dynamics (Wang, Jackson, and Zhuang, 2010; van der Elst, 2017). The majority of earthquakes occur in a small fraction of space and time near other earthquakes, and there are laws, such as the Omori-Utsu power-law, that describe the decay of aftershocks and seismic intensity (Utsu & Ogata, 1995). In "Perspectives on Clustering and Declustering of Earthquakes", Zaliapin and Ben-Zion (2021) present a reliable and straightforward measure, G, to analyze the clustering of earthquakes in space and time. The goal of this measure is to separate the effects of earthquakes concentrated around a complex fault network from other linked fluctuations in space and time. The authors suggest that the vast majority of earthquakes in the catalog are due to uneven spatial distribution, which conceals signals from combined space-time fluctuations. To improve the accuracy of the catalog, they recommend various techniques for detecting and removing these combined fluctuations. The metrics for declustering the catalog include the goals

of separating the signals from earthquakes, aftershocks, foreshocks, swarms, and other types of clusters to examine or remove them from the catalog.

The ROC-based Gini coefficient is used to measure the degree of coupled space-time clustering, which is shown to be stronger than what is suggested by visual inspections and ETAS modeling. The article suggests that catalog declustering should be done before further analysis to uncover additional features of seismicity beyond the strong clustering caused by aftershocks. The ROC diagram provides a convenient assessment of the coupled space-time clustering, and a large Gini coefficient value indicates a concentration of events in a small fraction of the examined space-time volume. The article proposes that the quality of declustering should be assessed by removing catalog inhomogeneities and biases, rather than focusing on the final product. The declustering procedure used in the article tries to remove clustered events and allows the user to decide which events are background or clustered. The article notes that the factorized rate may include temporal variations that are not related to event-event triggering, and that alternative declustering techniques can result in different numbers of background events.

Declustering is not the main focus of this work, so it is left to the reader to learn more about the topic and the aforementioned paper is a great place to start (Zaliapin and Ben-Zion, 2020).

3 Background

At the heart of developing a statistical model is determining how the processes being modeled are correlated (Gelman & Hill, 2006). There are many different ways of mathematically and statistically representing the correlation between two things, and this section will provide a review of different ways to quantify the relationship between two quantities. After this review, we will dive deeply into one such method, the ROC-based Gini coefficient introduced in the previous section, and apply it to the field of seismology discussed above.

3.1 Comparing Measures

To study the relationship between quantities we first establish properties of the quantities themselves. Studying measures generalizes the intuitive notions of length, area, and volume. Let X be a set such that $\mathrm{X} \subset \mathbf{R}^n$. For example, X is a space in two-dimensions. Usually, we consider up to three-dimensions for our space. The power set of X denoted $P(X)$ is then the set of all subsets of X. For example, if X is the set of two elements, a and b, the $X = a, b$ and the power set of X is

$$P(X) = \{\emptyset, \{a, b\}, \{a\}, \{b\}\}. \tag{22}$$

A measurable set is defined as any subset A of the power set of X, $P(X)$, such that $A \subseteq P(X)$

The axioms of probability for a measure f are as follows:

If f is a measure if $\forall A, B \subset X$,

$$0 \leq f(A) \leq 1 \tag{23}$$

The measure f is said to be normalized if the total probability,

$$f(X) = 1 \tag{24}$$

and

$$f(A) + f(B) = f(A \cup B) \; if \; A \cup B = \emptyset \tag{25}$$

This can be extended to n and infinite events.

A counting measure is a type of measure used to assign a non-negative value to a set, indicating its ”size” or ”cardinality”. Let (X, Σ) be a measurable space. The measure f on X defined by

$$f(A) = \begin{cases} n & \text{if } A\text{has exactly}n \text{ elements} \\ \infty & \text{otherwise} \end{cases} \tag{26}$$

for all $A \in \Sigma$ is called the counting measure on X.

Intuitively, a counting measure assigns a non-negative integer value to each element of the measurable space, and the measure of a set is simply the sum of the measures of its elements. Counting measures are used in probability theory to define discrete probability distributions, where the probability of an event is proportional to its "size" or frequency. While a counting measure can be defined on an infinite set, in practice it is typically used to measure finite subsets of the set (Taylor, 2018).

A level set of a real-valued function f of n real variables is a set where the function takes on a given constant value c, that is:

$$L_c(f) = \{(x_1, \ldots, x_n) \mid f(x_1, \ldots, x_n) = c\}, \tag{27}$$

(Osher & Fedkiw, 2003).When the number of independent variables is two, a level set is called a level curve, also known as contour line; so a level curve is the set of all real-valued solutions of an equation in two variables x_1 and x_2. When $n = 3$, a level set is called a level surface. A level surface is the set of all real-valued roots of an equation in three variables x_1, x_2 and x_3. For higher values of n, the level set is a level hypersurface, the set of all real-valued roots of an equation in $n > 3$ variables.

A sublevel set of f is a set of the form

$$L_c^-(f) = \{(x_1, \ldots, x_n) \mid f(x_1, \ldots, x_n) \leq c\} \tag{28}$$

is called a sublevel set of f (or, alternatively, a lower level set or trench of f)(Osher & Fedkiw, 2003). A strict sublevel set of f is

$$\{(x_1, \ldots, x_n) \mid f(x_1, \ldots, x_n) < c\} \tag{29}$$

(Osher & Fedkiw, 2003). Similarly, a superlevel set of f (or an upper level set of f) is defined as

$$L_c^+(f) = \{(x_1, \ldots, x_n) \mid f(x_1, \ldots, x_n) \geq c\} \tag{30}$$

And a strict superlevel set of f is

$$\{(x_1, \ldots, x_n) \mid f(x_1, \ldots, x_n) > c\} \tag{31}$$

3.2 Wasserstein Metric, Monge Problem, Kantorovich problem

The Wasserstein metric is one such alternative approach to comparing measures. The Wasserstein metric, also known as the Earth Mover's Distance (EMD), is a measure of the distance between two probability distributions (Villani, 2003). It is defined as the minimum amount of "work" or cost required to transform one distribution into the other, where the cost is defined as the amount of mass transported multiplied by the distance it is transported. It was first introduced by Gaspard Monge in 1781 in the context of optimization problems in transportation theory, known as the Monge problem. The Monge problem asks to find the most efficient way of transporting material from one pile to another pile, with the cost of transportation being proportional to the amount of material transported and the distance it is moved.

The Wasserstein metric formalizes this idea by considering the total cost of transportation, or the amount of "work" required, as the distance between two distributions (Villani, 2003). Given two probability distributions p and q, the Wasserstein distance between them is defined as the minimum cost of transforming one distribution into the other, where the cost is defined as the integral of the transportation cost function over all pairs of points in the two distributions. Formally, let (M, d) be a metric space that is a Radon space. For $p \in [1, \infty)$, the wasserstein p-distance between two probability measures μ and ν on M with finite p-moments is

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbf{E}_{(x,y)\sim\gamma} d(x,y)^p \right)^{\frac{1}{p}} \tag{32}$$

where $\Gamma(\mu,\nu)$ is the set of all couplings of μ and ν (Villani, 2003). A coupling γ is a joint probability measure on $M \times M$ whose marginals are μ and ν on the first and second factors, respectively. That is,

$$\int_M \gamma(x,y)dy = \mu(x) \tag{33}$$

$$\int_M \gamma(x,y)dx = \nu(y) \tag{34}$$

In the context of the Monge problem and the Wasserstein metric, the final integral is the expected cost of the transportation between two distributions, which is the minimum cost that is achieved when transforming one distribution into the other. Assume that there is given some cost function $c(x,y) \geq 0$ that gives the cost of transporting a unit mass from the point x to the point y, then the total cost of a transport plan γ is

$$\int\int c(x,y)\gamma(x,y)dxdy \tag{35}$$

The final integral provides a measure of the distance between the two distributions, with lower values indicating that the distributions are more similar.

The Monge problem is a special case of the more general Kantorovich problem, which involves finding the optimal transportation plan between two probability distributions. The Kantorovich problem allows for a more general class of cost functions, whereas the Monge problem is restricted to the case where the cost function is a distance function (Villani, 2009). In the Monge-Kantorovich problem, a cost function is a mathematical representation of the cost or distance between two objects, usually in the context of mass transportation. The cost function determines the cost of moving a unit of mass from one point to another. The goal of the Monge-Kantorovich problem is to

find an optimal transport plan that minimizes the total cost of transportation, which is typically defined as the sum of the cost function values over all unit masses transported from the source to the target.

If the Wasserstein distance between two measures is small, then one of the measures is potentially a localized version of the other. We will explore this concept of localization at length in the Methodology section.

4 Methodology

Statistically speaking, this method was developed for finding a relationship between several processes. In defining the relation between measures (f_i) the following questions were considered: (i) Are f_1 and f_2 related? (ii) How are they related? (iii) Can we use one of the measures to predict (or forecast) the other?

4.0.1 Measures and Level Sets

To begin to answer these questions we will set up the following general mathematically framework.

Consider a set X (which can be multidimensional) partitioned into a finite collection of m non-overlapping subsets

$$X = X_i : 1, ..., m. \tag{36}$$

A general measure f on a set X is a function $f : \Sigma \to [0, \infty]$, where Σ is a sigma-algebra on X, satisfying the following properties:

- $f(\emptyset) = 0$, where $\emptyset$ denotes the empty set.
- f is countably additive: for any sequence of pairwise disjoint sets $\{A_n\}$ in Σ, we have $f(n = 1^{\infty} A_n) = \Sigma_{n=1}^{\infty} f(A_n)$.

- f is translation-invariant: for any set A in Σ and any x in X, we have $f(A + x) = f(A)$, where $A + x = a + x : a \in A$.

where the density of measure f is

$$f(I) = \int_I f(x)dx \tag{37}$$

and we will assume that all measures have a density (Bauer, 2001).

An upper level set of a measure f will be defined as

$$L_f(c) = \{X_i : f(x) \geq c\} \tag{38}$$

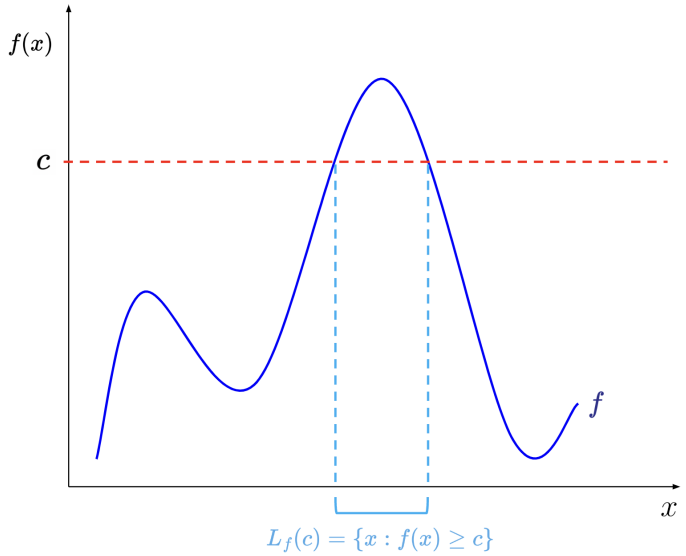

Figure 7: Level set of a measure f

Consider a pair of measures now f and g. To compare measures f and g, we can measure the level sets of one measure with respect to the other measure. The measure of the level set $L_f(c)$ with respect to g is

$$I_{f|g}(c) = \int_{X_i \subset L_f(c)} g(X_i)dx \tag{39}$$

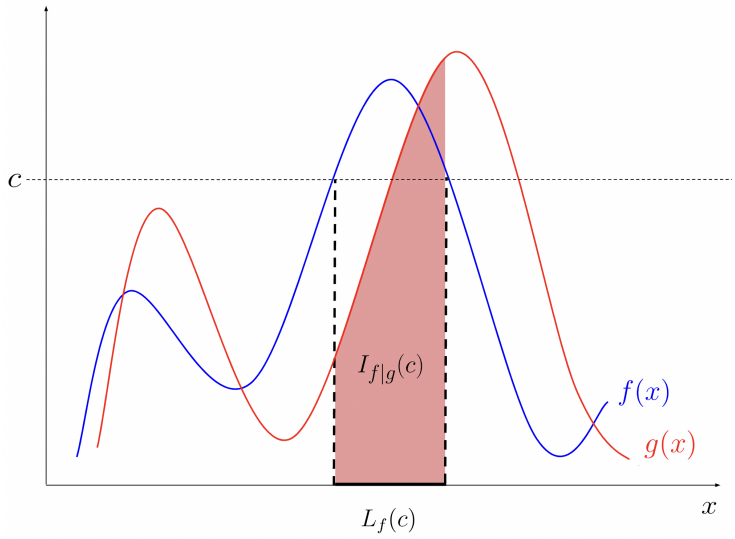

Figure 8: Mass $I_{g|f}(c)$ of the level set $L_f(c)$ with respect to measure $g(x)$ equals the area of the shaded region. Measure $f(x)$ is shown by the blue line and measure $g(x)$ is shown by the red line.

4.0.2 Receiver Operating Characteristic (ROC)

To compare measures on intervals, the Receiver Operating Characteristic (ROC) can be used. To define the ROC consider three measures f, g and h. The ROC of measure f with respect to the ordered pair of measures g and h, denoted by $R(f, g|h)$, is the set $(I_{f|g}(c), I_{f|h}(c))$ where parameter $c \geq 0$. Every level set $L_f(c)$ corresponds to a specific point in the set $R(f|g, h)$. The measure of the level set $L_f(c)$ corresponding to measures g and h is represented by the x and y coordinates of the point, respectively. The normalized (probability) measures used in this work adhere to the condition $f(X) = g(X) = h(X) = 1$. The following properties for the ROC can be inferred for normalized probability measures:

- The set $R(f|g, h)$ is contained within the unit square $[0, 1] \times [0, 1]$
- The points $(I_{f|g}(\infty), I_{f|h}(\infty)) = (0, 0)$ and $(I_{f|g}(0), I_{f|h}(0)) = (1, 1)$ are in the set $R(f|g, h)$.
- The set $R(f|g, h)$ consists of a finite number of points; the number of points is one more than the number of distinct values of $P(X_i), i = 1, ..., m$

- The set $R(f|g,h)$ is monotone non-decreasing: $I_{f|g}(a) < I_{f|g}(b)$ implies $I_{f|h}(a) < I_{f|h}(b)$ for any non-negative pair a, b.
- For any pair of measure f and g, the set $R(f|g,g)$ lie on the line connecting the points $(0,0)$ and $(1,1)$.
- For the uniform measure U that assigns the same value to each subset X_i, and any pair of measures g and h, the set $R(U|g,h)$ consists of the two points: $(0,0)$ and $(1,1)$.

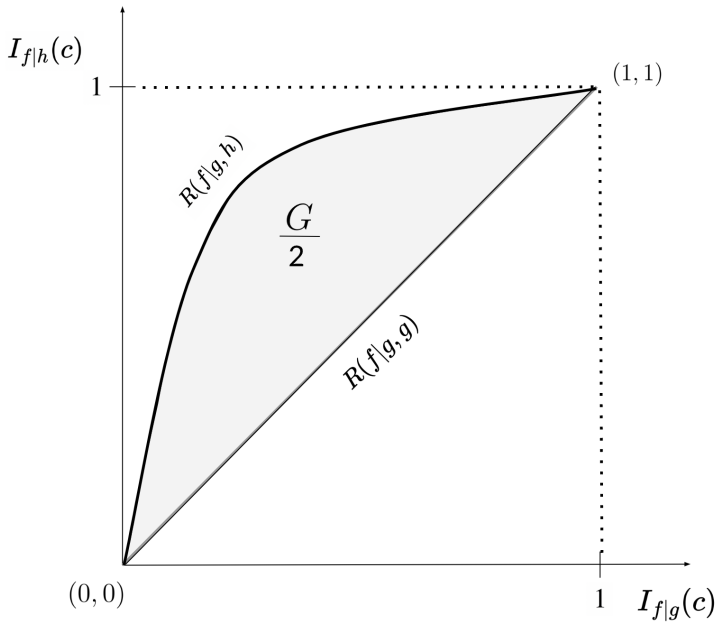

Figure 9: ROC diagram $R(f,g|h)$: a parametric plot of $I_{f|g}(c)$ vs $I_{f|h}(c)$ for $c \geq 0$. The set $R(f|g,g)$ always lies on the diagonal (grey).

4.0.3 Absolute Localization

The ROC diagram can be used to quantifying the degree of concentration of a given measure within a specific domain. To quantify the degree of concentration of a measure, compare the measure to a uniform measure using the following properties.

For the uniform measure U which assigns equal values to all subsets x_i, and any two measures f and g:

- The set $R(f|U,f)$ includes two points, namely (0,0) and (1,1).
- The set $R(f|U,f)$ is situated on or above the diagonal line that connects points (0,0) and (1,1).
- If $f \neq U$ then $R(f|U,f)$ encompasses at least one point above the diagonal line.

The absolute localization of measure f can be quantified as the amount of deviation of $R(U,f|f)$ from the diagonal line. A point (x,y) in the set $R(f|U,f)$ corresponds to a level set L where $f(L) = y$ and $U(L) = x$. The point (x,y) will be located close to the corner point $(0,1)$, far from the diagonal line, if a significant portion of f is concentrated within a small region of L.

The Gini coefficient can be used to provide a precise measure of absolute localization. The Gini coefficient for absolute localization, denoted G_f, is twice the area between the curve of $R(f|U,f)$ and the diagonal line. G_f can take on values between 0 and 1, where $G_f = 0$ indicates $f = U$, and G_f gets closer to 1 when f becomes more heterogeneous (more localized).

4.0.4 Relative Localization

Relative localization is a measure of the localization of measure f with respect to another non-uniform measure g. Relative localization means that measure f is concentrated within the same domain as g and it has more pronounced peaks (its measure of absolute localization is higher). This concepts is illustrated in Figure (10) where f(red) is a more localized version of g(blue), f has a larger absolute concentration than g and is located within the same domain.

To measure relative localization use the receiver operating characteristic for the set $R(g|f,g)$ with every point in the set corresponding to a particular level set $L_g(c)$. This set may include points below the diagonal line, indicating that the corresponding level sets of measure f accumulate a higher mass with respect to measure g than with respect to g itself. This situation is interpreted as relative localization of f with respect to g.

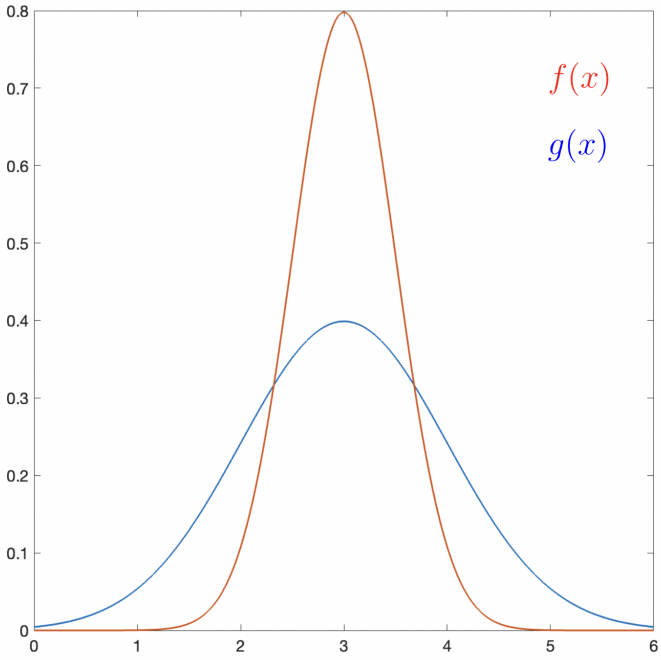

Figure 10: Measure f is a more localized version of measure g.

To quantify the relative localization of measure f with respect to measure g, we calculate the area above ROC curve $R(g|f,g)$, denoted as $G_{f|g}$. The parameter $G_{f|g}$ ranges from 0 to 1. If $G_{f|g} < \frac{1}{2}$ then measures f and g are not concentrated within the same regions. If f is a localized version of g then $G_{f|g} > \frac{1}{2}$

4.0.5 ROC and Random Variables

A general measure can be changed into a probability measure. A probability measure is a function P defined on a sigma-algebra Σ of subsets of a sample space S, such that:

- *Non-negativity:* For any set $A \in \Sigma, P(A) \leq 0$.
- *Normalization:* $P(S) = 1$.
- *Countable additivity:* For any sequence of disjoint sets $A_1, A_2, ..., A_n \in \Sigma$, the probability of their union is the sum of their individual probabilities:

$$P(A_1 U A_2 U ... U A_n) = P(A_1) + P(A_2) + ... + P(A_n). \tag{40}$$

The sigma-algebra Σ defines the collection of events for which probabilities can be assigned, and the axioms ensure that the probabilities are non-negative, normalized, and additive.

To change a measure f to a probability measure v, you need to normalize the measure by dividing each set's measure by the measure of the entire space. This normalization ensures that the measure of the entire space is equal to 1. Define the new measure v on Σ as follows:

$$v(A) = f(A)/f(X) \tag{41}$$

for all A in Σ. v must satisfy the properties of a measure: non-negativity, countable additivity, and $v(\emptyset) = 0$.

4.0.6 Theoretical ROC for a Random Variable

Let W be a normalized random variable with probability density function(PDF) $f_W(w)$ and the cumulative distribution function (CDF) $F_W(w)$. The ROC diagram of W with respect to a uniform random variable U, denoted $R(W|U, W)$ can be obtained by plotting $I_{U|W}(w_0) = P(W > w_0)$ on the x-axis and $I_{W|W}(w_0) = \int_{w_0}^{\infty} wf(w)dw$ on the y-axis. The equation of a theoretical ROC curve for the random variable, can then be parameterized as follows:

$$x = P(W > w_0) = 1 - F(w_0) \tag{42}$$

and

$$y = \int_{w_0}^{\infty} wf(w)dw. \tag{43}$$

The theoretical ROC curve for a random variable can sometimes be parameterized in terms of just x and y and will be a monotonically increasing curve that starts at $(0, 0)$ and ends at $(1, 1)$ with equation $R(W|U, W)$. The Gini coefficient for a random variable

is thus:

$$G = 2\int_0^1 R(W|U,W) - xdx \tag{44}$$

4.0.7 Theoretical ROC for an Exponential Random Variable

Let W be a normalized exponential random variable with parameter $\lambda = 1$. The PDF of W is

$$f_W(w) = e^{-w} \quad \text{if } w \geq 0 \tag{45}$$

and the CDF for W is

$$F_W(w) = 1 - e^{-w} \quad \text{if } w \geq 0. \tag{46}$$

The coordinates of the ROC curve can be parameterized as follows:

$$x = e^{-w_0} \tag{47}$$

$$y = (w_0 + 1)e^{-w_0} \tag{48}$$

Solving Equation (47) for w_0, yields

$$w_0 = ln(x) \tag{49}$$

Substituting this in for w_0 in Equation (48) yields

$$y = -xlnx + x \tag{50}$$

The theoretical ROC curve for an exponential random variable is therefore:

$$R(W|U,W) = -xlnx + x \quad \text{for } 0 < x \leq 1 \tag{51}$$

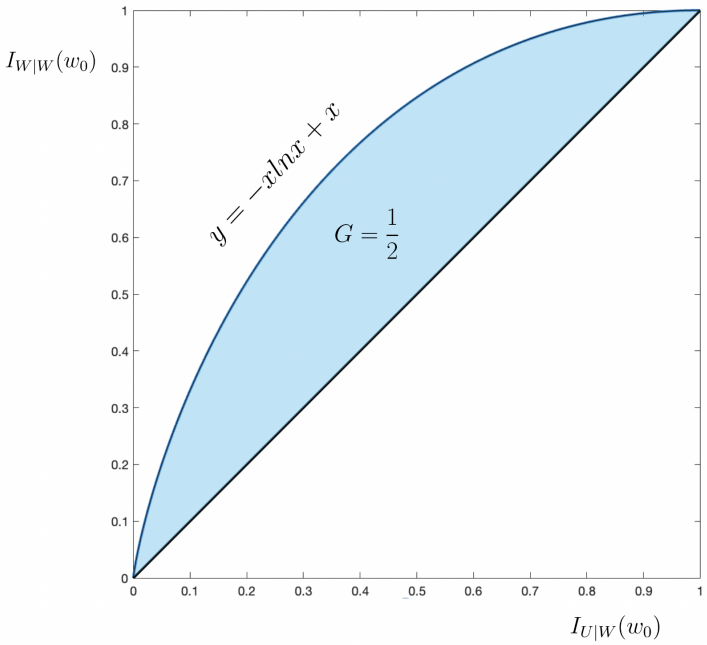

Figure 11: ROC diagram $R(W|U,W)$: a parametric plot of $I_{U|W}(w_0)$ vs $I_{W|W}(w_0)$ for $w_0 \geq 0$.

The Gini coefficient for an exponential variable is thus:

$$G = 2\int_0^1 -xlnxdx = \frac{1}{2} \tag{52}$$

4.0.8 Theoretical ROC for a Pareto Random Variable

Let W be a normalized pareto random variable with parameters $w_m = 1$ and $\alpha \geq 1$. The PDF of W is

$$f_W(w) = \frac{\alpha w_m^\alpha}{w^{\alpha+1}}, w \geq w_m, \tag{53}$$

and CDF

$$F_W(w) = 1 - \left(\frac{w_m}{w}\right)^{-\alpha}, w \geq w_m. \tag{54}$$

Normalizing this probability distribution by its mean $(E(w) = \frac{\alpha w_m}{\alpha - 1})$ yields the following PDF and CDF

$$f_W(w) = \frac{\alpha(\frac{\alpha-1}{\alpha})^\alpha}{w^{\alpha+1}} \tag{55}$$

$$F_W(w) = 1 - \left(\frac{\alpha-1}{\alpha x}\right)^\alpha \tag{56}$$

The x-coordinates of the ROC for a normalized pareto random variable can be parameterized as follows:

$$x = \left(\frac{\alpha-1}{\alpha w_0}\right)^\alpha \tag{57}$$

and the y-coordinates can be parameterized:

$$y = \left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1} w_0^{1-\alpha}. \tag{58}$$

Solving Equation (57) for w_0, yields:

$$w_0 = \left(\frac{\alpha-1}{\alpha}\right) x^{-\frac{1}{\alpha}}. \tag{59}$$

Substituting this in for w_0 in Equation (58) yields:

$$y = x^{\frac{\alpha-1}{\alpha}}. \tag{60}$$

The theoretical ROC curve for an Pareto random variable will therefore be a monotonically increasing curve that starts at $(0, 0)$ and ends at $(1, 1)$ and has equation:

$$R(W|U, W) = x^{\frac{\alpha-1}{\alpha}}, 0 < x \leq 1, \alpha > 1 \tag{61}$$

The Gini coefficient for an Pareto variable is thus:

$$G = 2\int_0^1 \left(x^{\frac{\alpha-1}{\alpha}} - x\right)dx = \frac{1}{2\alpha-1} \tag{62}$$

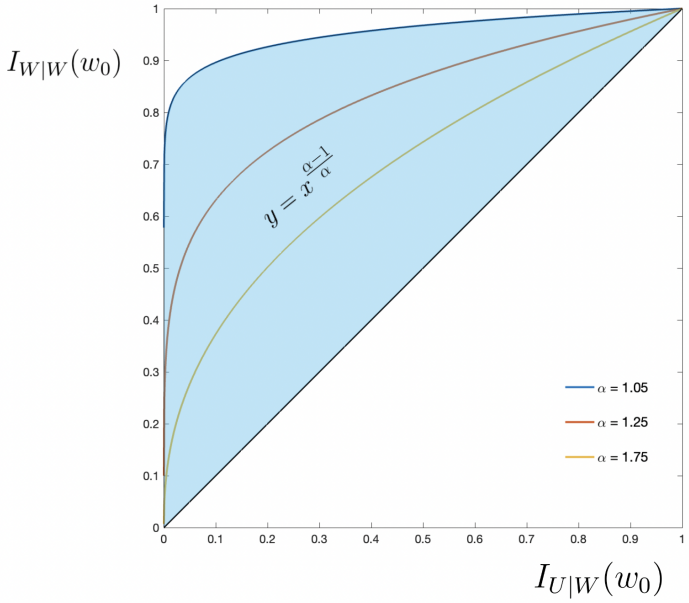

Figure 12: ROC diagram $R(W|U,W)$: a parametric plot of $I_{U|W}(w_0)$ vs $I_{W|W}(w_0)$ for $w_0 \geq 0$ for a Pareto random variable with parameter $\alpha = 3$.

4.0.9 Theoretical ROC for a Bivariate Random Variables

Let (W, Z) be continuous random variable that happen to be correlated with joint probability density function(PDF) $f_{W,Z}(w, z)$ and the cumulative distribution function (CDF) $F_{W,Z}(w, z)$. Consider a subset $L_W(w_0)$ corresponding to a threshold on the random variable W such that $W > w_0$.

The ROC diagram of (W, Z) with respect to W, denoted $R(Z|W, Z)$, is the set $R(Z|W, Z) = \{(x, y) : y = f(x)\}$. To obtain the theoretical ROC curve of Z with respect to W, the x-coordinates, normalized by the mean $E[W] = \int_0^\infty w f_W(w)dw$ where $f_W(w) = \int_0^\infty f(w, z)dz$, can be parameterized as follows:

$$x = I_{Z|W}(w_0) = \frac{\int_{w_0}^\infty \int_{-\infty}^\infty z f_{W,Z}(w, z)dzdw}{\int_{-\infty}^\infty w f_W(w)dw} \tag{63}$$

and the y-coordinates, also normalized by the mean, can be parameterized

$$y = I_{W|W}(w_0) = \frac{\int_{w_0}^\infty w f_W(w)dw}{\int_{-\infty}^\infty w f_W(w)dw} \tag{64}$$

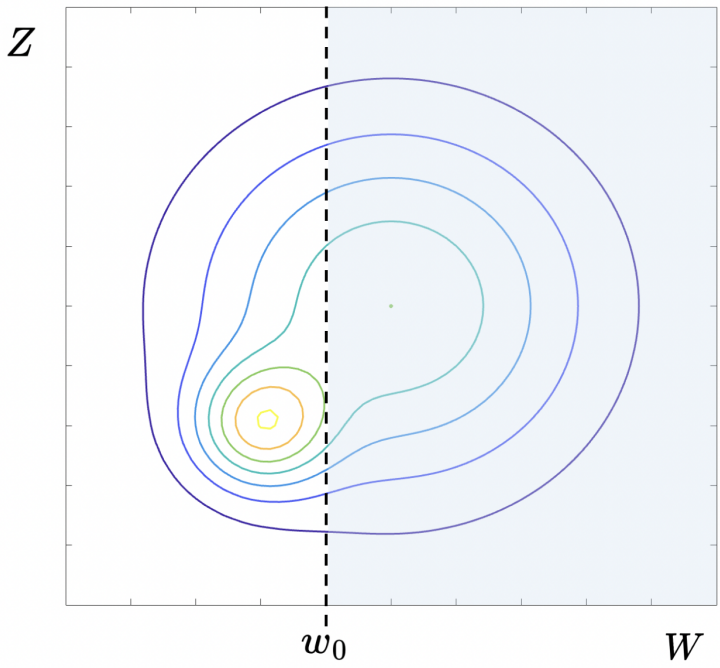

Figure 13: A level set $L_W(w_0)$ corresponding to a threshold on the random variable W such that $W > w_0$.

Like the theoretical ROC curve for a random variable, the ROC for a two random variables can sometimes be parameterized in terms of just x and y and will be a monotonically increasing curve that starts at $(0, 0)$ and ends at $(1, 1)$. The Gini coefficient will be defined by Equation (44).

4.0.10 Theoretical ROC for a Bivariate Exponential Random Variables

Let (W, Z) be bivariate exponential random variable with joint probability density function(PDF)

$$f_{W,Z}(w, z) = e^{-(\lambda_1 w + \lambda_2 z + \theta w z)}\big[(\lambda_2 + \theta w)(\lambda_1 + \theta z) - \theta\big]. \tag{65}$$

where $w, z > 0$ and $\lambda_1, \lambda_2, \theta > 0$. The cumulative distribution function (CDF)

$$F_{W,Z}(w, z) = e^{-(\lambda_1 w + \lambda_2 z + \theta w z)} \tag{66}$$

.

For the bivariate exponential distribution, the mean we will normalize by is

$$E[W] = \int_0^\infty w f_W(w) dw = \frac{1}{\lambda_1} \tag{67}$$

The x-coordinate of the ROC for a bivariate exponential random variables can be parameterized as follows:

$$x = \frac{\int_{w_0}^\infty \int_0^\infty z e^{-(\lambda_1 w + \lambda_2 z + \theta w z)} \big[(\lambda_2 + \theta w)(\lambda_1 + \theta z) - \theta\big] dz dw}{\frac{1}{\lambda_1}} \tag{68}$$

$$= \lambda_1 \int_{w_0}^\infty \frac{(\lambda_1 \theta w + \theta + \lambda_1 \lambda_2) e^{-\lambda_1 w}}{(\theta w + \lambda_2)^2} dw \tag{69}$$

$$= \lambda_1 \big[\frac{-e^{-\lambda_1 w}}{(\lambda_2 + \theta w)} |_{w_0}^\infty\big] \tag{70}$$

$$= \frac{\lambda_1}{(\lambda_2 + \theta w_0) e^{\lambda_1 w_0}} \tag{71}$$

and the y-coordinate, normalized by the mean, is:

$$y = \frac{\int_{w_0}^\infty w \lambda_1 e^{-\lambda_1 w} dw}{\frac{1}{\lambda_1}} \tag{72}$$

$$= \lambda_1 \big[-\frac{(aw + 1)\, \mathrm{e}^{-aw}}{a} |_{w_0}^\infty \tag{73}$$

$$= \lambda_1 \frac{(\lambda_1 w_0 + 1)\, \mathrm{e}^{-\lambda_1 w_0}}{\lambda_1} \tag{74}$$

$$= (\lambda_1 w_0 + 1)\, \mathrm{e}^{-\lambda_1 w_0} \tag{75}$$

Neither of these coordinates can be solve for w_0 so the final parameterized coordinates for the ROC diagram for a bivariate exponential random variables is:

$$ROC(Z|W, W) = \{(x, y) : \big(\frac{\lambda_1}{(\lambda_2 + \theta w_0) e^{\lambda_1 w_0}}, \frac{\lambda_1 w_0 + 1}{e^{\lambda_1 w_0}}\big)\} \tag{76}$$

where $w_0 \in [0, \infty)$ and $\lambda_1, \lambda_2, \theta > 0$.

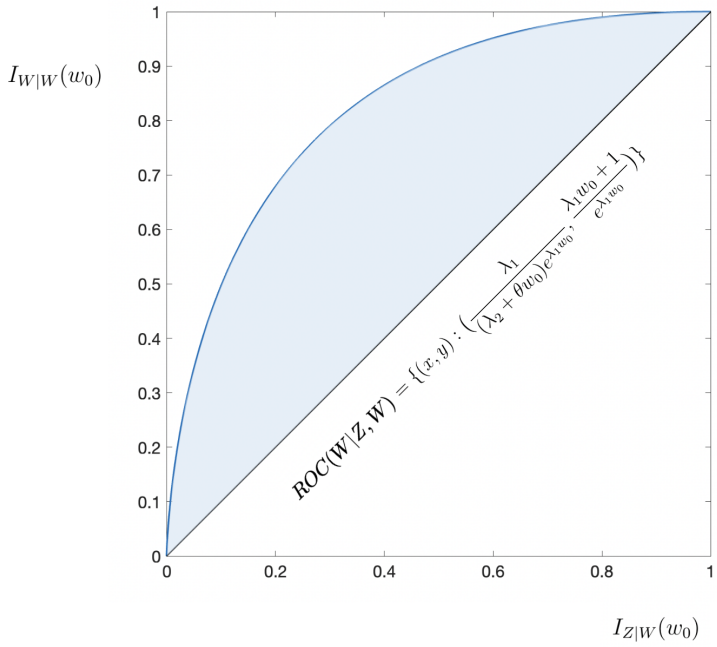

Figure 14: ROC diagram for bivariate exponential random variables with $\lambda_1, \lambda_2, \theta = 2$ and $w_0 \in [0, \infty)$.

4.0.11 Simulation

The theoretical results for the bivariate exponential ROC diagram presented above where verified by a numerical simulations using the model of Balakrishna and Shijia (2014).

To evaluate the theoretical procedure discussed above we carried out a simulation for different sample sizes and for different values of the parameters. As outlined by Balakrishna and Shijia (2014), for the simulation, we first generate realizations from positive stable RV's $\{U_i\}$ using the relation:

$$U = E^{-\frac{1-\alpha}{\alpha}} (\sin\xi)^{\frac{-1}{\alpha}} \sin(\alpha\xi) \sin((1-\alpha)\xi)^{\frac{1-\alpha}{\alpha}}, \tag{77}$$

that was first proposed by McKenzie (1982), where ξ is a uniform RV over $(0, \pi)$ and E is a unit exponential RV independent of ξ . For specified values of the parameters we simulated independent and identically distributed (i.i.d.) sequence $\{w_i\}$ from Exponential(λ) distribution and then obtained the sequence $\{z_i\}$ using the relation

$$z_i = w_i^{\alpha} \frac{1}{\beta} \left(\frac{\lambda}{U_i}\right)^{\alpha}, i = 1, 2, ...n \tag{78}$$

This generates a bivariate sample $\{w_i, z_i\}$. From this sample we can generate the ROC diagram and Gini coefficient for the absolute and relative localization as outlined above. In Figure (15) you can see an example of the results of the simulation.

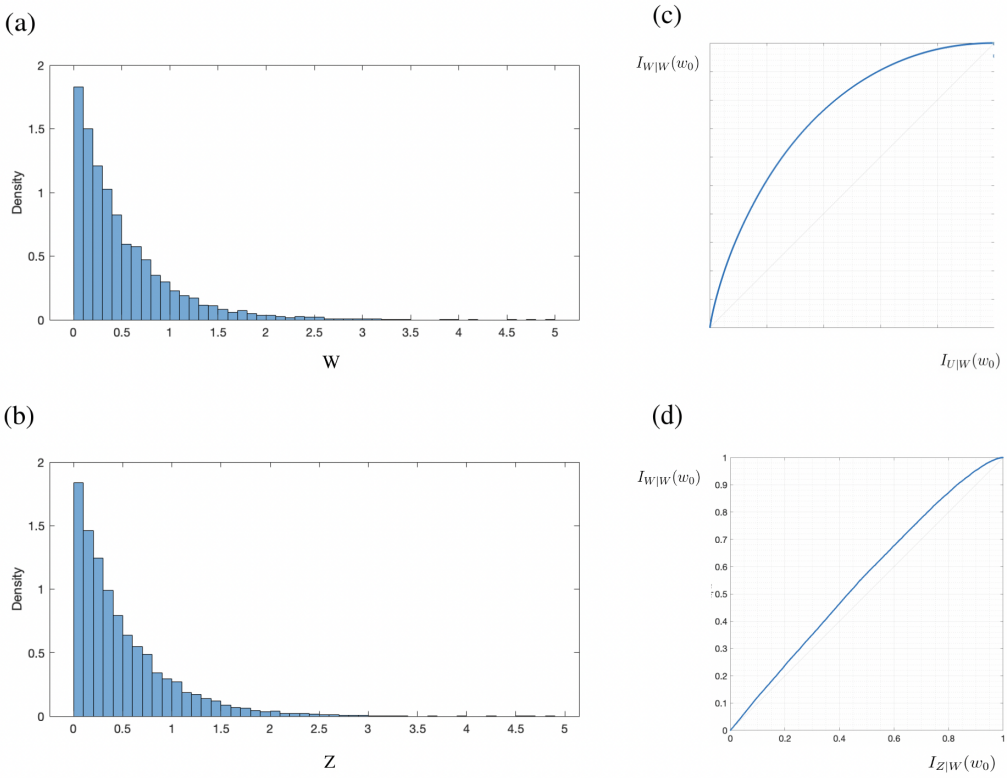

Figure 15: Results for a bivariate exponential simulation with parameters $\alpha = 0.3$, $\beta = 2.0$, $\lambda = 0.5$ for samples of size $n = 500$ with 100 replicates. (a) Marginal distribution of simulated random variable W. (b) Marginal Distribution of simulated random variable Z. (c) $ROC(W|U, W)$ where $G = 0.46$. (d) $ROC(W|Z, W)$ where $G = 0.446$.

This simulation can be used to test the performance of the Gini coefficient as compared to the Pearson correlation coefficient. To do this, the simulation was carried out for different sample sizes and for different specified values of the parameters. Then the values of the relative Gini coefficient and the Pearson correlation coefficient for between the simulated bivariate exponential (W, Z) were computed. Results for this simulation are displayed in the Table (1).

n	α	β	λ	Gini Coefficient		Pearson Correlation	
				Mean	SD	Mean	SD
50	0.8	2.0	2.0	0.4500	0.0278	0.7985	0.0871
	0.7	2.0	3.0	0.4271	0.0300	0.7076	0.1002
	0.6	3.0	1.0	0.4048	0.0300	0.5998	0.1130
	0.5	1.0	2.0	0.3774	0.0393	0.4985	0.1406
	0.3	2.0	0.5	0.3285	0.0424	0.3098	0.1410
	0.2	0.5	2.0	0.3037	0.0492	0.1936	0.1579
100	0.8	2.0	2.0	0.4473	0.0191	0.7903	0.0635
	0.7	2.0	3.0	0.4251	0.0213	0.6986	0.0812
	0.6	3.0	1.0	0.4010	0.0259	0.5999	0.0949
	0.5	1.0	2.0	0.3756	0.0282	0.5094	0.1055
	0.3	2.0	0.5	0.3264	0.0268	0.3120	0.1066
	0.2	0.5	2.0	0.2996	0.0242	0.1906	0.1015
500	0.8	2.0	2.0	0.4501	0.0077	0.8025	0.0314
	0.7	2.0	3.0	0.4248	0.0091	0.6977	0.0345
	0.6	3.0	1.0	0.4248	0.0091	0.6977	0.0345
	0.5	1.0	2.0	0.3766	0.0121	0.4970	0.0481
	0.3	2.0	0.5	0.3264	0.0130	0.3049	0.0519
	0.2	0.5	2.0	0.3007	0.0139	0.2008	0.0515

Table 1: Simulation results for the Bivariate Exponential Random Variables W and Z for different values of the parameters. 100 replicates of each sample where created of the given sample size n for each of the sets of parameter values in the table.

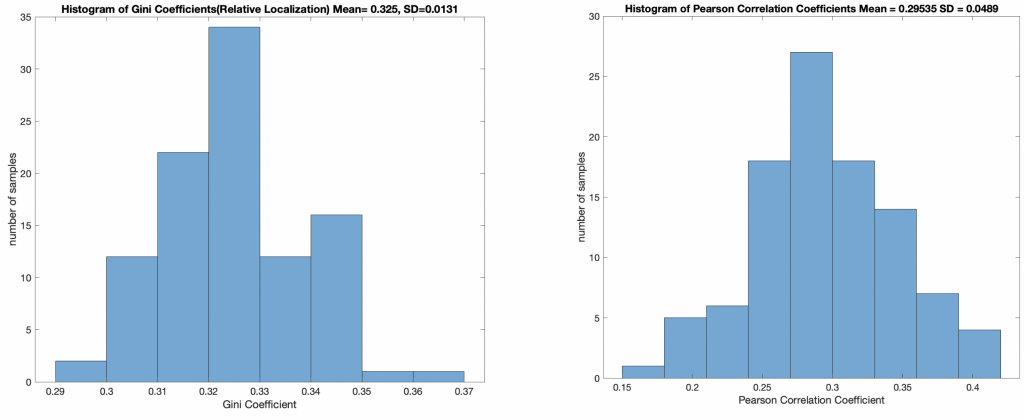

Figure 16: Comparing the Gini Coefficient to Pearson Correlation Coefficient of 100 samples of size $n = 500$ for parameters $\alpha = 0.3$, $\beta = 2.0$, and $\lambda = 0.5$.

When comparing the Gini coefficient to the Pearson correlation coefficient, as seen in Figure (16), we see that the Gini Coefficient is a more stable measure. When conduct the simulation for 100 replicates of different sample size with different parameters, the Gini coefficient consistently produces sampling distribution with a smaller standard deviation.

4.0.12 Independence

Theorem: When$W$and Z are independent, $ROC(Z|W, W)$ will be the same as $ROC(W|U, W)$.

Proof: Let (W, Z) be continuous normalized random variable that are independent. Then the joint probability density function(PDF) $f_{W,Z}(w, z) = f_W(w)f_z(z)$ and the cumulative distribution function (CDF) $F_{W,Z}(w, z) = F_W(w)F_Z(z)$. Consider a subset $L_W(w_0)$ corresponding to a threshold on the random variable W such that $W > w_0$.

The theoretical ROC curve of Z with respect to W, can be parameterized as follows:

$$x = I_{Z|W}(w_0) = \frac{\int_{w_0}^{\infty}\int_{-\infty}^{\infty} z f_{W,Z}(w, z)dzdw}{\int_{-\infty}^{\infty} w f_W(w)dw} \tag{79}$$

$$= \frac{\int_{w_0}^{\infty}\int_{-\infty}^{\infty} z f_W(w)f_z(z)dzdw}{\int_{-\infty}^{\infty} w f_W(w)dw} \tag{80}$$

$$= \frac{E[Z]\int_{w_0}^{\infty} f_W(w)dw}{E[W]} \tag{81}$$

$$= (1 - F_W(w_0)) \tag{82}$$

$$= I_{U|W}(w_0) \tag{83}$$

which is the x-coordinate of $R(W|U, W)$.

Recall the y-coordinates, also normalized by the mean, can be parameterized

$$y = I_{W|W}(w_0) = \int_{w_0}^{\infty} w f_W(w)dw \tag{84}$$

which is exactly the same parameterization of the y-coordinate for the $R(W|U, W)$.

5 Results - Localization of Seismicity

The above methodology can be used as a novel statistical approach to quantify regional earthquake clustering. Clustering was introduced in the background section as a fundamental component of understanding seismicity and earthquake triggering mechanisms. Using the ROC diagram to measure space-time clustering allows disentangling effects related to concentration of events around a heterogeneous regional fault network from coupled space-time fluctuations.

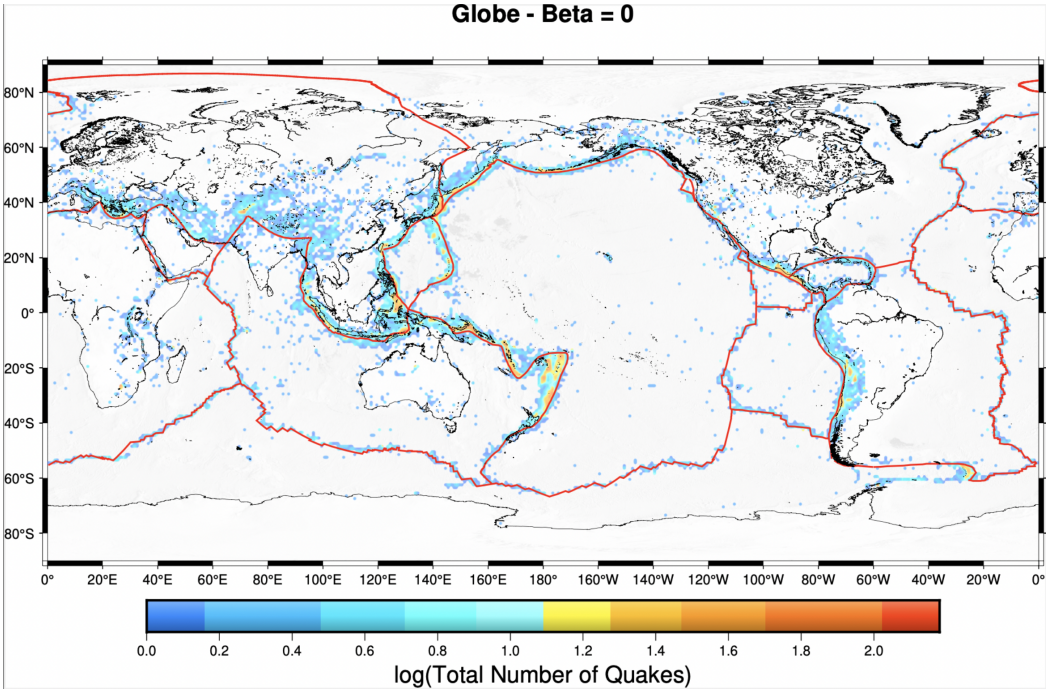

Figure 17: Clustering of global seismicity based on the ComCat catalog during the period of 1981–2022. Shades of color reflect the number of earthquakes (logarithmic scale; see color bar) with $M \geq 4.5$ in square spatial cells with side length equal to 1°. Notice a significant space inhomogeneity.

In this section the ROC will be used to examine and illustrate seismic clustering in multiple seismically active regions, including the Reno area. The ROC will be used to explore several general measures of seismic rate that can account for the number of events, and the total area of fault breaks. This analysis will systematically examine general and coupled space-time clustering of raw and declustered catalogs. The Gini coefficient will be presented as a simple and robust measure of space-time clustering to

illustrate and quantify earthquake clustering with examples of seismicity from regions worldwide.

5.1 Data

For this study, the ANSS comprehensive Earthquake Catalog was used. The regions examined for this study are Reno, southern California, Japan, New Zealand, Italy, the Atlantic and Pacific. Specifics of the analysis for each region are in Table (2).

5.2 Methodology

A systematic assessment of the inhomogeneity of the space-time distribution of seismicity in the study regions was conducted using ROC diagrams, which produce a single measure of space-time clustering, the Gini coefficient (G). As mentioned in the previous section, the coefficient G may assume values between 0 and 1, a value of G close to 1 in this case indicates a large portion of events are concentrated in a small fraction of the examined space-time volume.

More formally, we partition the examined space-time area into voxels with space dimension Δ and time size ω. The voxels can be indexed by two space indices i, j and time index k. Measure seismic activity within a voxel at location $x = i, j$ at time t by the equation

$$\Sigma(x,t) = \sum_{i=1}^{k} 10^{\beta M_i} \tag{85}$$

where the summation is taken over all events, $1, ..., k$, within the voxel, β is a parameter we change to understand the different characteristics of clustering in that region, and M_i is the magnitude of event i. $\beta = 0$ corresponds to counting events, $\beta = 1$ approximates the faultbreak area, and $\beta = \frac{3}{2}$ corresponds to the seismic moment. We only examine voxels with the time-integrated value of Σ being larger than a threshold Σ_0 (in this work, $\Sigma_0 = 0$).

Region	Space Range	Time Range	Magnitude Range	Number of Events	Size of Space Resolution, Δ ($^\circ$)	Size of Temporal Resolution, w (years)
SoCal	$124^\circ - 114^\circ$W $30^\circ - 38^\circ$N	$1971 - 2022$	$2.3 - 7.3$	113,290	0.6	1.5
Japan	$128^\circ - 146^\circ$E $30^\circ - 47^\circ$N	$1971 - 2022$	$4.5 - 9.1$	15,754	1	5
New Zealand	$160^\circ - 180^\circ$E $50^\circ - 30^\circ$S	$1990 - 2022$	$4.4 - 7.8$	3,153	1	6
Italy	$0^\circ - 25^\circ$E $30^\circ - 50^\circ$N	$1983 - 2005$	$3.3 - 6$	4,811	1	6
Reno	$121^\circ - 117^\circ$W $38^\circ - 41^\circ$N	$2000 - 2022$	$2.5 - 6.5$	3,937	0.6	5
Atlantic	55°N.$W - 5^\circ$E 60°S $- 60^\circ$N	$1970 - 2010$	$4.5 - 7.1$	4,551	5	5
Pacific	$80^\circ - 170^\circ$E; 25°S $- 20^\circ$N	$1973 - 2010$	$4.5 - 9.1$	57,024	2.5	1

Table 2: Parameters of ComCat Catalog Analysis. Threshold $\Sigma_0 = 0$ for all regions.

After computing Σ for each non-empty voxel, sort the voxels by there Σ value in descending order: $\Sigma_1 \geq \Sigma_2 ... \geq \Sigma_m$, assuming there are m non-empty voxels. First, we evaluate the general clustering where the ROC diagram is a plot of Σ within the most active voxels (y-axis) vs. the fraction of the examined non-empty voxels (x-axis). For example, if $\beta = 0$, we are simply counting the number of events in each non empty voxel, $\Sigma_i = N_i$, and the ROC diagram essential consists of the points

$$\left(\frac{i}{m}, \frac{N_1 + ... + N_i}{N}\right) \text{ for } i = 1, ..., m, \tag{86}$$

where N is the total number of examined events: $N = \sum_{i=1}^{m} N_i$.

Next, in order to remove the effects of marginal space and time heterogeneities in the examined catalog, we evaluate coupled space-time clustering. In this analysis, the ROC diagram is a plot of Σ within the most active voxels (y-axis) vs. the weighted fraction of the examined non-empty voxels(x-axis). The weights are determined by the factorized space-time rates of background events. Declustering is done by the method of Zaliapin and Ben-Zion. The x-axis of the factorized ROC diagram is scaled in a way that represents the product of the estimated marginal space and time rates of background seismicity , denoted as $J(x,t) = S(x)T(t)$, corresponding to the diagonal. Essentially, the x-axis measures the proportion of the factorized rate $J(x,t)$ in the most active cells of the process under study, while the y-axis shows the proportion of events in the process. Any deviation from the diagonal is due to coupled space-time irregularities, such as short-term fluctuations in local areas. The Gini coefficient G of this scaled ROC diagram measures the overall level of coupled space-time clustering.

The ROC diagram for a factorized rate, which is scaled to account for possible marginal space and time catalog inhomogeneities is defined in the following way. Start by estimating the background field using the method of Zaliapin and Ben-Zion (2020) and counting the numbers $B(i,j,k)$ of estimated background events in the space-time voxels indexed by $\{i,j,t\}$. The marginal space and time background rates are estimated as

$$S(i,j)) = \sum_k B(i,j,k) \text{ and } T(k) = \sum_i \sum_j B(i,j,k), \tag{87}$$

where the space variable x is discretized by grid $\{i,j\}$ and the time variable t is discretized by grid k. The factorized rate J on the voxel grid indexed by $\{i,j,k\}$ is defined as

$$J(i,j,k) = \frac{J_0(i,j,k)}{\sum_{i,j,k} J_0(i,j,k)}, \text{ with } J_0(i,j,k) = S(i,j)T(k) \tag{88}$$

To do this for a region, we sort the non-zero $\Sigma(i,j,k)$ value for each voxel in descending order: $\Sigma_1 \geq \Sigma_2 ... \geq \Sigma_m$. This creates the order map $n(i,j,k,t) : \{i,j,k\} \rightarrow \{n\}$ that maps the space-time indices $\{i,j,k\}$ and order index $n = 1, \dots m$. We use the inverse map $\{i(n), j(n), k(n)\} : \{n\} \rightarrow \{i,j,k\}$ to form a sequence of factorized counts:

$$J_n = J(i(n), j(n), k(n)), n = 1, \dots m, \tag{89}$$

where the 1-st factorized count is taken from the voxel with the maximal value of Σ, Σ_1 of the observed events, the 2-nd count is taken from the voxel with the second largest count Σ_2 of the observed events, etc. The x-axis in the scaled ROC diagram shows the proportion of the factorized rate $J(i,h,k)$ within the most active cells of the examined process, and the y-axis shows the observed proportion of events in the same cells. In other words, the diagram consists of the points

$$\big(J_i, \frac{(\Sigma_1 + \dots + \Sigma_n)}{\Sigma_N}\big) \text{ for } i = 1, \dots, m, \tag{90}$$

where $\Sigma_N = \sum_{i=1}^m \Sigma_i$ is the total value of Σ in the full catalog.

Factorized space-time rates of background events are a way to model the expected frequency of events occurring in a particular region over a certain period of time. In seismology, this is often used to model the frequency of earthquakes in a specific area over a given time period. The term "factorized" refers to the fact that the model separates the

rate of events into two components: space and time. The space component represents the expected frequency of events in different locations, while the time component represents the expected frequency of events at different times. The model assumes that the rate of events is proportional to the product of the space and time components. In other words, the expected frequency of events in a given region at a given time is the product of the expected frequency of events in that region and the expected frequency of events at that time.

Figure (18) illustrates the general procedure for quantifying clustering of earthquakes in southern California with the ROC diagram. The first plot is a measure of earthquake clustering with respect to a uniform measure. The red shading corresponds to the most active voxels in our region over the period of time for the catalog. Clearly seismicity in the southern California area is not uniform over the region, instead, a large fraction of earthquakes occurs in a relatively small time-space volume. We call this "clustering" and measure it by a general Gini coefficient.

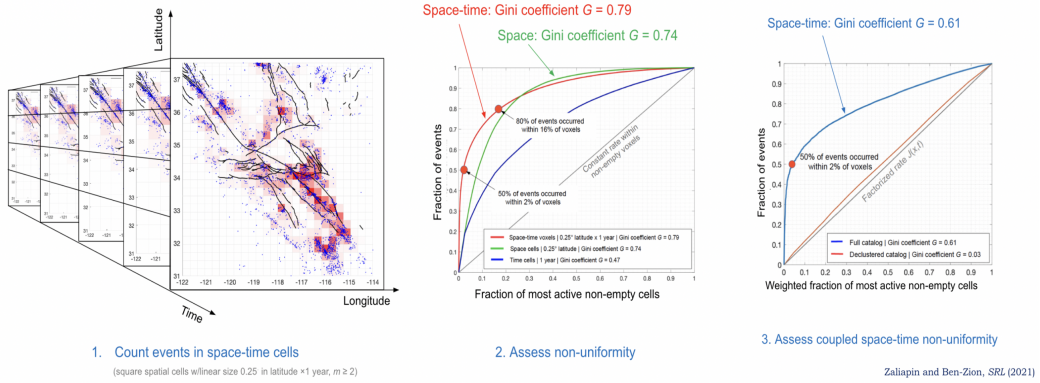

Figure 18: Quantifying clustering of earthquakes in southern California with the receiver operating characteristic (ROC) diagram. The analysis uses the catalog of Hauksson et al. (2012, extended) during 1981–2020 with magnitude $M \geq 2$, and space–time voxels with square space projection of latitude size $0.25v$ and 1 yr duration.

5.3 Results

Figures of the above implement method for all regions of this study, except the Atlantic and Pacific, are located in Appendix A to help the reader visualize clustering in each

Catalog	Clustering with Respect to Constant Rate, G	Background Clustering with Respect to Constant Rate, G	Clustering with Respect to Factorized Rate, G	Background Clustering with Respect to Factorized Rate, G
SoCal	0.79	0.53	0.55	0.07
Japan	0.71	0.57	0.35	0.06
New Zealand	0.62	0.50	0.28	0.03
Italy	0.64	0.42	0.62	0.05
Reno	0.78	0.42	0.22	0.12
Atlantic	0.49	0.53	0.26	0.04
Mean	0.67	0.49	0.37	0.06
Standard Deviation	0.10	0.07	0.15	0.03

Table 3: Localization Results ($\beta = 0$). The ComCat catalog was used for all regions.

regions. This Appendix contains maps of clustered and declustered seismic activity in each region, as well as the ROC diagrams for both clustering with respect to a constant rate and clustering with respect to a factorized rate. This section summarizes the numeric findings of the analysis.

5.3.1 $\beta = 0$

The results displayed in Figure (19) show the Gini coefficients G for all regions for $\beta = 0$.

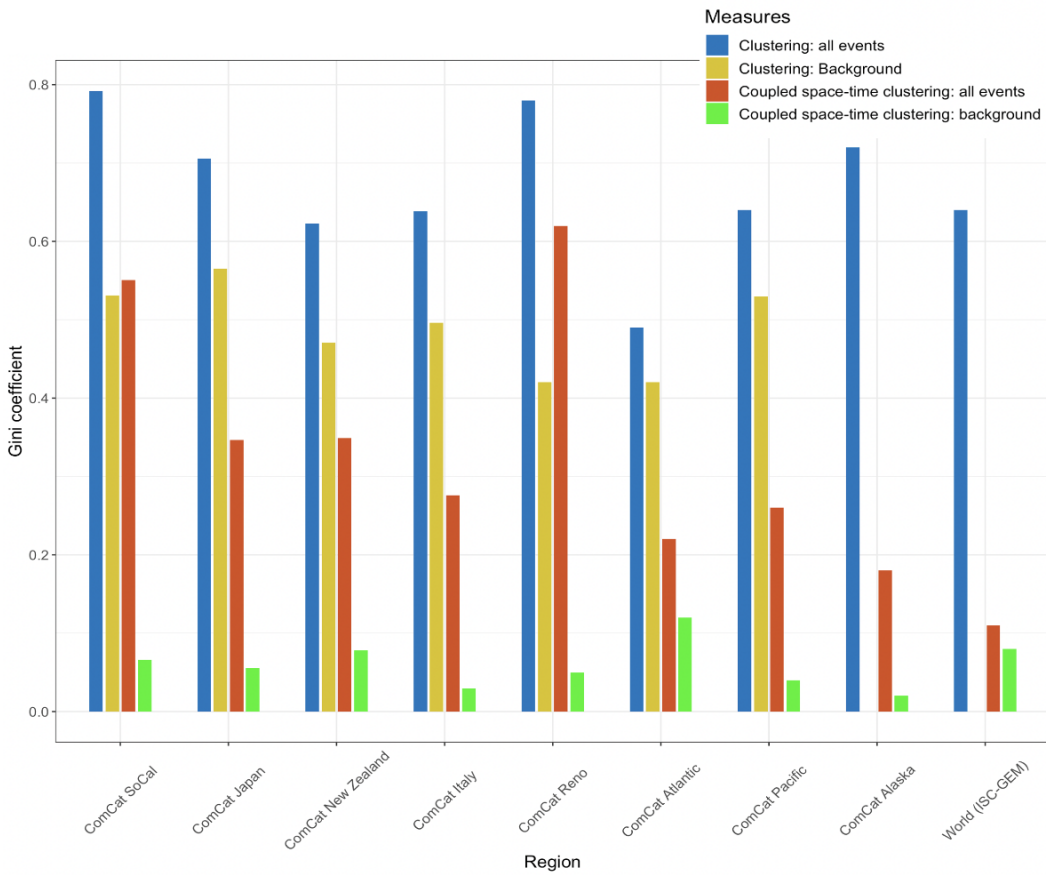

Figure 19: Summary of G for $\beta = 0$ by region.

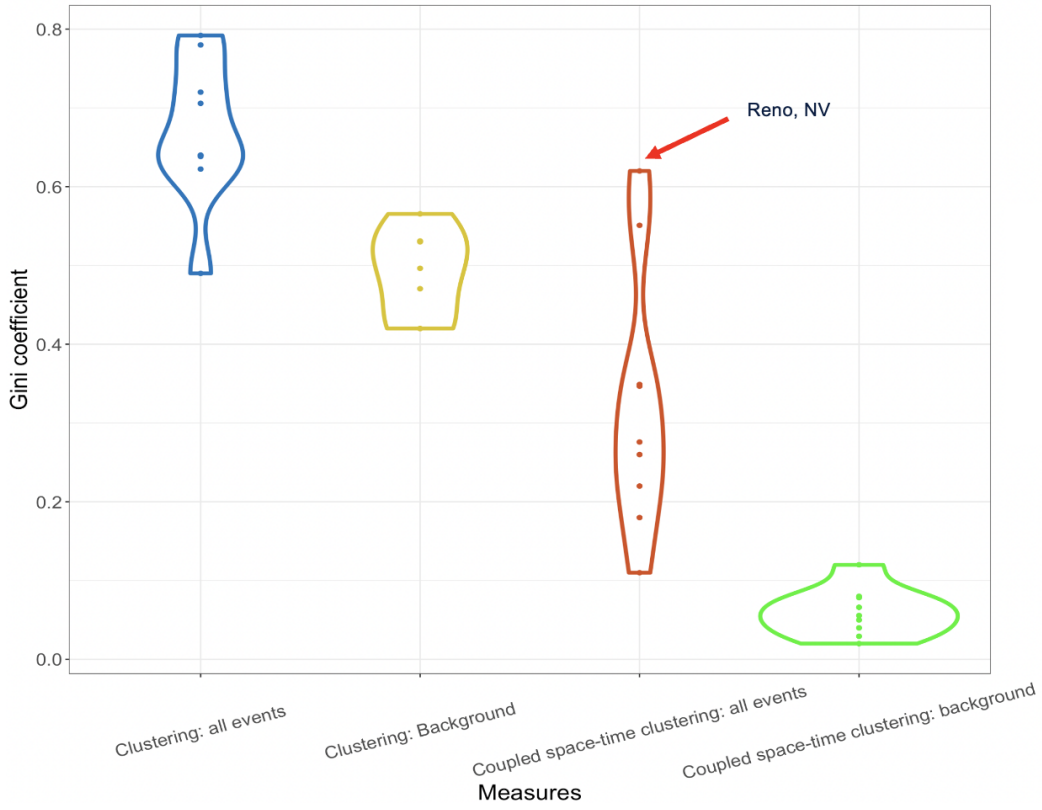

Figure 20: Summary of G for $\beta = 0$ by measure.

5.3.2 $\beta = \frac{1}{2}$

Adjusting the parameter in Equation (85) to $\beta = \frac{1}{2}$ we get the following results displayed in Table (4) illustrated in Figures (21) and (22).

Catalog	Clustering with Respect to Constant Rate, G	Background Clustering with Respect to Constant Rate, G	Clustering with Respect to Factorized Rate, G	Background Clustering with Respect to Factorized Rate, G
SoCal	0.80	0.55	0.57	0.14
Japan	0.71	0.58	0.36	0.08
New Zealand	0.64	0.51	0.38	0.15
Italy	0.65	0.51	0.30	0.06
Reno	0.80	0.51	0.61	0.12
Atlantic	0.49	0.44	0.19	0.12
Pacific	0.66	0.56	0.30	0.10
Mean	0.68	0.52	0.39	0.11
Standard Deviation	0.11	0.05	0.15	0.03

Table 4: Localization Results ($\beta = \frac{1}{2}$). The ComCat catalog was used for all regions.

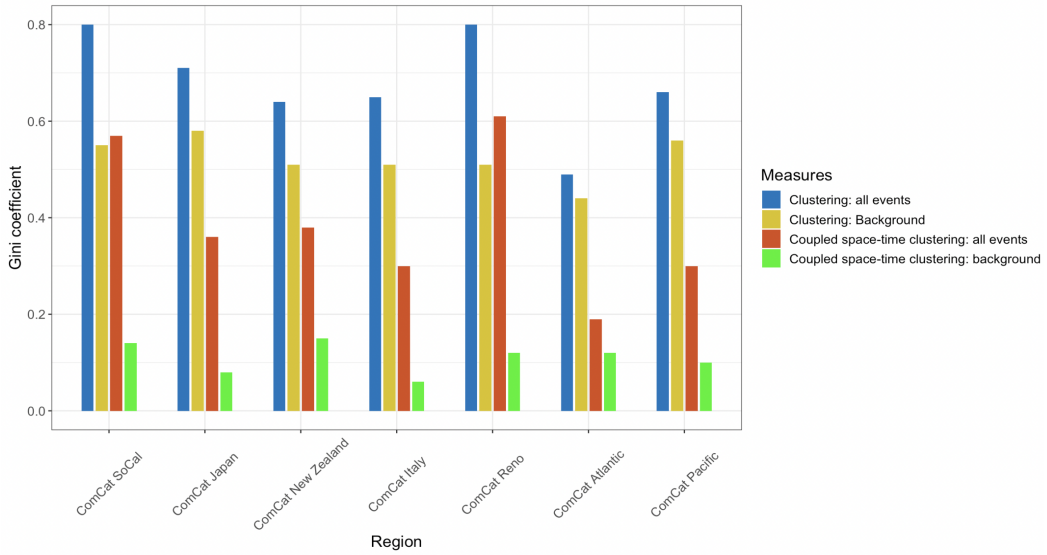

Figure 21: Summary of G for $\beta = \frac{1}{2}$ by region.

Catalog	Clustering with Respect to Constant Rate, G	Background Clustering with Respect to Constant Rate, G	Clustering with Respect to Factorized Rate, G	Background Clustering with Respect to Factorized Rate, G
SoCal	0.91	0.86	0.75	0.70
Japan	0.87	0.77	0.64	0.44
New Zealand	0.80	0.83	0.55	0.61
Italy	0.73	0.66	0.40	0.24
Reno	0.89	0.85	0.73	0.66
Atlantic	0.61	0.62	0.22	0.26
Pacific	0.82	081	0.61	0.61
Mean	0.80	0.77	0.56	0.50
Standard Deviation	0.11	0.10	0.19	0.19

Table 5: Localization Results $(\beta = 1)$. The ComCat catalog was used for all regions.

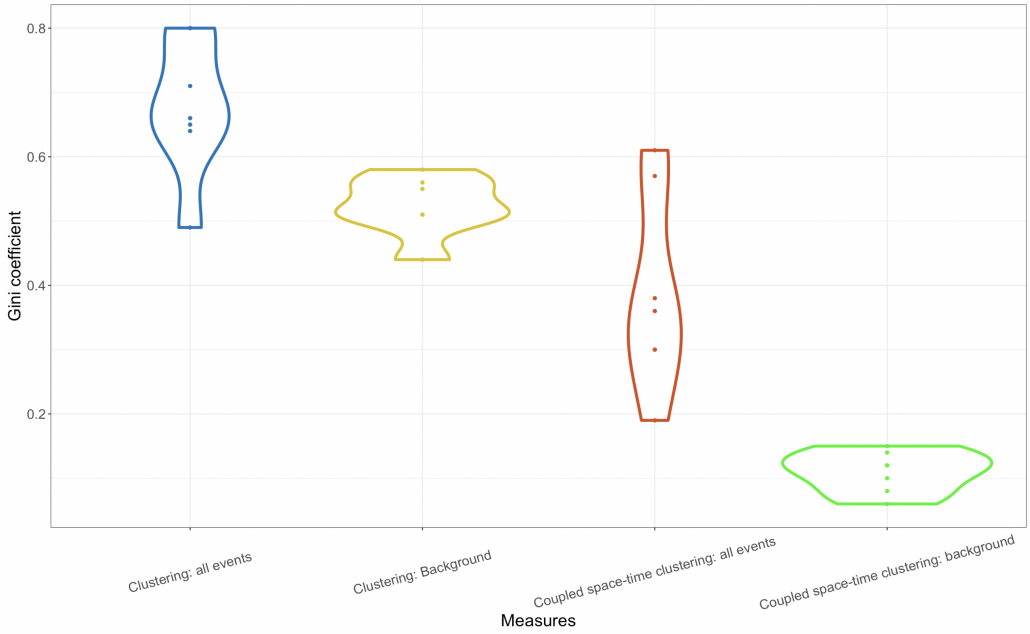

Figure 22: Summary of G for $\beta = \frac{1}{2}$ by measure.

5.3.3 $\beta = 1$

Changing the parameter in Equation (85) to $\beta = 1$ we get the following results displayed in Table (5) illustrated in Figures (23) and (24).

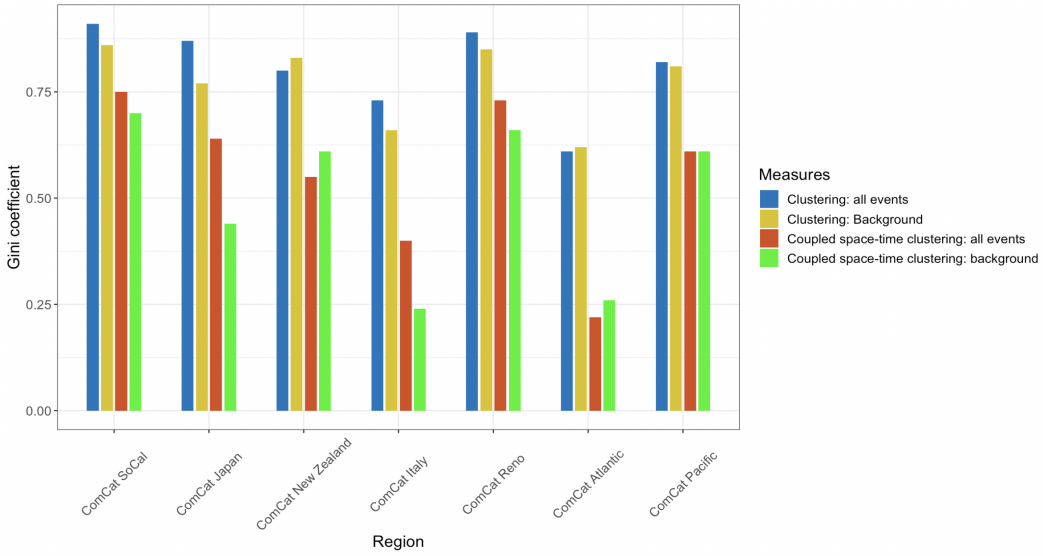

Figure 23: Summary of G for $\beta = 1$ by region.

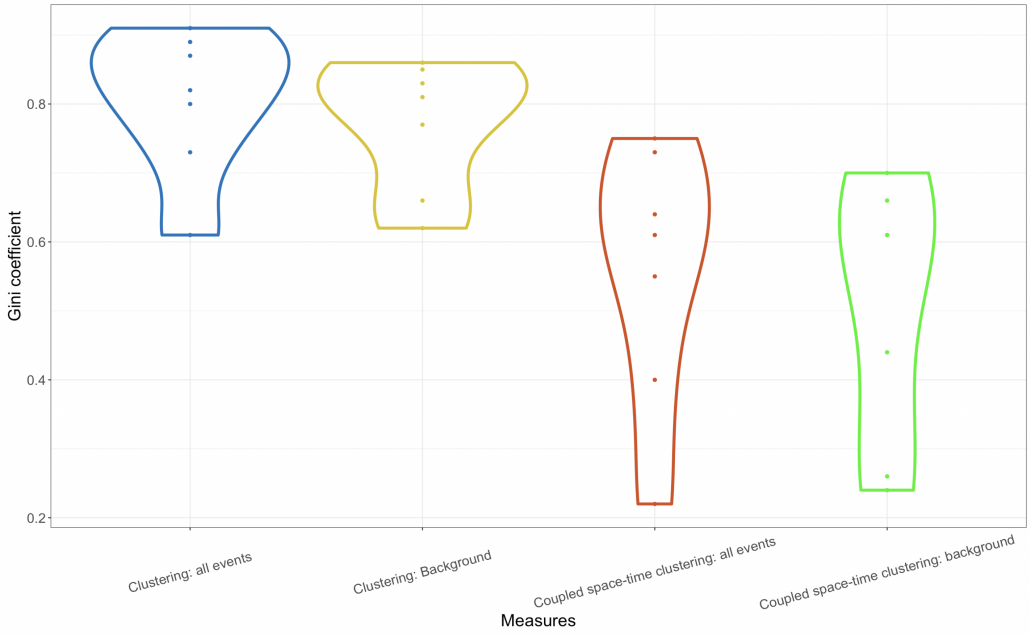

Figure 24: Summary of G for $\beta = 1$ by measure.

5.4 Conclusions

This study examined clustering in regions worlds wide and found that overall clustering is high for both the raw and declustered catalogs for all values of β. Even when we are not considering empty voxels, there are not the same number of earthquakes in

each non-empty voxel. Coupled space-time clustering had the largest variation region to region and therefore provides the most insight into region specific triggering mechanisms and causes of clustering. This is supported by the fact that the variability (as measured by standard deviation) for this measure is the largest. This suggests that the marginal space clustering, fault network, plays a dominant role in the overall clustering.

An interesting observation gained from Figure (19)is that this coupled clustering (the red measure) is rather strong, and varies substantially from catalog to catalog. For the other measures, clustering for all events, for background events and coupled space time clustering for background events (the blue yellow and green measures in the figures) we see a more uniform clustering over all regions. The Gini coefficient for coupled space-time clustering for all events can help us identify regional differences in clustering because there is the greatest variability in this measure. Coupled clustering of declustered catalogs (for $\beta = 0$) is negligible ($G < 0.1$).

This violin plot grouping Gini coefficients by measure for $\beta = 0$ instead of by region highlights this fact that coupled space-time clustering for all events has the largest variability of all the measures. The region with the largest coupled space time clustering was Reno, NV.

Results for $\beta = \frac{1}{2}$ lead to similar conclusions as the results for $\beta = 0$. Coupled space-time clustering for all events has the largest variability region to region of all the measures suggesting that regional differences in triggering mechanisms and characteristics of siesmisity are best captured by this measure.

Results for $\beta = 1$ show far less variability in measures region to region than the results for $\beta = 1$ or $\beta = \frac{1}{2}$. General clustering for all events is very similar to general clustering for background events when $\beta = 1$. Coupled space-time clustering for all events is also similar to that of background events.

In conclusion, the Gini coefficient for coupled space-time clustering can be used as a metric that shows variability in clustering region to region.

6 Discussion

Earthquake prediction is a difficult task due to the complex and not fully understood dynamics involved (Turcotte 1997; Keilis-Borok 2002; Ben-Zion 2008; Jordan et al. 2011). This thesis contributions to the field of seismology by providing a method of visualizing, quantifying and comparing spatial-temporal clustering region to region.

First we measure earthquake clustering with respect to a uniform distribution of seismicity. Seismicity in all regions is clearly not uniform over the region, and instead, a large fraction of earthquakes occurs in a relatively small time-space volume. We call this ”clustering” and measure it by a general Gini coefficient. This general clustering is attributed to two reasons: first, earthquakes occur along fault zones (see Figure (17)). Second, there exist groups of events highly localized in a given area during a short time (these would be aftershock sequences after a large earthquake). We would like to separate these two effects. Technically, this is done by introducing the space-time coupling Gini coefficient.

In other words, we start by using a measure that shows no clustering if events occur uniformly in space and time. Then we introduce a more delicate measure that doesn't include clustering related to events around hot spot areas that are active all the time, or burst of activity in time that might affect an entire region. Then we measure how strongly the observed clustering deviates from this new (factorized) model. All deviations represent burst of activity that are localized in time and space.

Factorized space-time rates of background events are a way to model the expected frequency of events occurring in a particular region over a certain period of time. In seismology, this is often used to model the frequency of earthquakes in a specific area over a given time period. The term ”factorized” refers to the fact that the model separates the rate of events into two components: space and time. The space component represents the expected frequency of events in different locations, while the time component represents the expected frequency of events at different times. The model assumes that the rate of events is proportional to the product of the space and time components. In other

words, the expected frequency of events in a given region at a given time is the product of the expected frequency of events in that region and the expected frequency of events at that time. Factorized space-time rates of background events are useful for estimating the probability of rare events, such as large earthquakes, occurring in a particular region over a given time period. By modeling the expected frequency of events based on historical data, seismologists can better understand the likelihood of future earthquakes and improve earthquake hazard assessments.

A next step for this research is to identify specifically what clustering mechanisms and physical characteristics are associated with regions that have a particularly low coupled space time clustering for all events or a particularly high Gini coefficient. This method could also be used to quantify clustering in other regions and for smaller regions within the areas of this study. Specifically, smaller regions of southern California focusing on specific faults. This method also has the potential to predict future seismicity from the observed long term distribution of seismicity.

Most importantly, this work introduces a novel approach for quantifying non-linear correlations between positive-valued processes and random variables. The method proposed uses the Receiver Operating Characteristic and the Gini coefficient to forecast one measure using another measure. We review the existing approaches to the problem both from a statistical perspective, reviewing other methods of determining if two variables are correlated and from a seismological perspective reviewing methods of quantifying clustering and modeling seismic activity. The theoretical advantage of our approach is that the ROC and Gini coefficients are directly related to the theoretical forecast problem. Practically, this method is easy to compute and implement in a variety of contexts. We expand the correlation technique to the time-space domain setting and adapt it to quantifying the degree of clustering in the observed earthquakes.

The Gini coefficient and the Pearson correlation coefficient are both measures of statistical dependence. Comparison with Pearson correlation for bivariate exponential random variables shows that the Gini coefficient is more stable (has smaller variability)

than the Pearson correlation coefficient for bivariate exponential random variables over many different values of the distribution parameters. Next steps for this research are to define the ROC diagrams for the Weibull distribution, either theoretically or empirically. Determining confidence intervals for the ROC curve for the bivariate exponential random variables would also allow for inference to be performed on ROC curves in applications. This method can be used in a broader context to quantify how systems produce extreme events. In the future, I hope this methodology could be applied in other areas outside seismology.

References

[1] Abercrombie, R. E. (2019). Similar starts for small and large earthquakes. Nature, 573, 42–43.

[2] ANSS: Advanced National Seismic System Composite Catalog. (2015). Available at: http://www.quake.geo.berkeley.edu/anss/.

[3] Babakus, E., & Yavas, U. (2008). Measurement of the effects of nonlinear relationships among variables. Journal of Business Research, 61(12), 1261-1271. https://doi.org/10.1016/j.jbusres.2008.01.012

[4] Baiesi, M., & Paczuski, M. (2004). Scale-free networks of earthquakes and aftershocks. Physical Review E, 69, 066106. doi:10.1103/PhysRevE.69.066106.

[5] Balakrishna, N., & Shiji, K. (2014). On a class of bivariate exponential distributions. Statistics & Probability Letters, 85, 153–160. https://doi.org/10.1016/j.spl.2013.11.009.

[6] Båth, M. (1965). Lateral inhomogeneities in the upper mantle. Tectonophysics, 2, 483–514.

[7] Bauer, H. (2001). Measure and integration theory. Berlin: de Gruyter.

[8] Bautista, J. (2011). Nearest-neighbor analysis of marked point fields with applications to cluster analysis of seismicity. Undergraduate Honors Thesis, Department of Mathematics and Statistics, University of Nevada, Reno.

[9] Ben-Zion, Y. (2008). Collective behavior of earthquakes and faults. Reviews of Geophysics, 46, RG4006.

[10] Ben-Zion, Y., Ampuero, J.-P. (2009). Seismic radiation from regions sustaining material damage. Geophysical Journal International, 178, 1351–1356. https://doi.org/10.1111/j.1365-246X.2009.04285.x.

[11] Ben-Zion, Y., Eneva, M., & Liu, Y. (2003). Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. Journal of Geophysical Research: Solid Earth, 108, 2307.

[12] Ben-Zion, Y., & Zaliapin, I. (2019). Spatial variations of rock damage production by earthquakes in southern California. Earth and Planetary Science Letters, 512, 184-193. https://doi.org/10.1016/j.epsl.2019.02.006.

[13] Bird, P., Liu, Z., & Rucker, W.K. (2008). Stresses that drive the plates from below: Definitions, computational path, model optimization, and error analysis. Journal of Geophysical Research, 113, B11406. doi:10.1029/2007JB005460.

[14] Bouchon, M., Durand, V., Marsan, D., Karabulut, H., & Schmittbuhl, J. (2013). The long precursory phase of most large interplate earthquakes. Nature Geoscience, 6, 299–302.

[15] Brillinger, D. (1988). Some statistical methods for random process data from seismology and neurophysiology. The 1983 Wold Memorial Lectures. Annals of Statistics, 16, 1–54.

[16] Brodsky, E. E., & Lajoie, L. J. (2013). Anthropogenic seismicity rates and operational parameters at the Salton Sea Geothermal Field. Science, 341(6145), 543–546.

[17] Casella, G., & Berger, R. L. (2002). Statistical inference (2nd ed.). Duxbury Press.

[18] Chen, X., Shearer, P.M. & Abercrombie, R.E. (2012). Spatial migration of earthquakes within seismic clusters in Southern California: Evidence for fluid diffusion, Journal of Geophysical Research: Solid Earth, 117, B04301. doi:10.1029/2011JB008973.

[19] Chu, A., Schoenberg, F., Bird, P., Jackson, D., & Kagan, Y. (2011). Comparison of ETAS parameter estimates across different global tectonic zones. Bulletin of the Seismological Society of America, 101(5), 2323–2339.

[20] Console, R., Murru, M., & Catalli, F. (2006). Physical and stochastic models of earthquake clustering. Tectonophysics, 417(1-2), 141–153.

[21] Daley, D. J., & Vere-Jones, D. (1972). A summary of the theory of point processes. In P. A. W. Lewis (Ed.), Stochastic Point Processes: Statistical Analysis, Theory and Applications (pp. 47-71). Wiley.

[22] Davidsen, J., Gu, C., & Baiesi, M. (2015). Generalized Omori-Utsu law for aftershock sequences in southern California. Geophysical Journal International, 201(2), 965–978.

[23] Dieterich, J. H. (1992). Earthquake nucleation on faults with rate-and state-dependent strength. Tectonophysics, 211(1-4), 115–134.

[24] Dieterich, J. (1994). A constitutive law for rate of earthquake production and its application to earthquake clustering. Journal of Geophysical Research, 99(B2), 2601–2618.

[25] Dresen, G., & Guéguen, Y. (2004). Damage and rock physical properties. In Y. Guéguen & M. Boutéca (Eds.), Mechanics of Fluid Saturated Rocks (International Geophysics Series, pp. 169–217). Elsevier.

[26] Eaton, D.W., Davidsen, J., Pedersen, P.K., & Boroumand, N. (2014). Breakdown of the Gutenberg-Richter relation for microearthquakes induced by hydraulic fracturing: influence of stratabound fractures. Geophysical Prospecting, 62(4), 806-818.

[27] Ellsworth, W.L. (2013). Injection-induced earthquakes. Science, 341(6142), doi:10.1126/science.1225942

[28] Ellsworth, W.L., & Bulut, F. (2018). Nucleation of the 1999 Izmit earthquake by a triggered cascade of foreshocks. Nature Geoscience, 11, 531-535.

[29] Embrechts, P., McNeil, A., & Straumann, D. (2013). Correlation and dependence in risk management: properties and pitfalls. In Risk management: Value at risk and beyond (pp. 176-223). Springer.

[30] Enescu, B., & Ito, K. (2001). Some premonitory phenomena of the 1995 Hyogo-Ken Nanbu (Kobe) earthquake: seismicity, b-value and fractal dimension. Tectonophysics, 338(3), 297-314.

[31] Eshelby, J.D. (1957). The determination of the elastic field of an ellipsoidal inclusion and related problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 241, 376-396.

[32] Felzer, K.R. (2008). Calculating California seismicity rates. In The Uniform California Earthquake Rupture Forecast, version 2 (UCERF 2), Appendix I, U.S. Geological Survey Open-File Report 2007-1437I and California Geological Survey Special Report 203I (pp. 42).

[33] Felzer, K.R., & Brodsky, E.E. (2006). Decay of aftershock density with distance indicates triggering by dynamic stress. Nature, 441(7094), 735-738.

[34] Freedman, D.A. (2005). Statistical models: Theory and practice (p. 246). Cambridge University Press.

[35] Gardner, J.K., & Knopoff, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Possionian? Bulletin of the Seismological Society of America, 64(5), 1363-1367.

[36] Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

[37] Genest, C., & Favre, A.C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering, 12(4), 347-368.

[38] Gutenberg, B., & Richter, C. (1954). Seismicity of the earth and associated phenomena (Second). Princeton University Press.

[39] Gu, C., Schumann, A.Y., Baiesi, M., & Davidsen, J. (2013). Triggering cascades and statistical properties of aftershocks. Journal of Geophysical Research, 118(8), 4278-4295.

[40] Hainzl, S. (2013). Comment on "Self-similar earthquake triggering, Bath's law, and foreshock/aftershock magnitudes: Simulations, theory, and results for southern California" by P.M. Shearer, J. geophys. Res.: Solid Earth, 118, 1188–1191.

[41] Hainzl, S., Moradpour, J., & Davidsen, J. (2014). Static stress triggering explains the empirical aftershock distance decay. Geophysical Research Letters, 41(24), 8818–8824.

[42] Hamiel, Y., Liu, Y., Lyakhovsky, V., Ben-Zion, Y., & Lockner, D. (2004). A viscoelastic damage model with applications to stable and unstable fracturing. Geophysical Journal International, 159, 1155–1165. https://doi.org/10.1111/j.1365-246X.2004.02452.x.

[43] Hanks, T. C., & Kanamori, H. (1979). A moment magnitude scale. Journal of Geophysical Research, 84(B5), 2348–2350.

[44] Harte, D. (1998). Dimension estimates of earthquake epicenters and hypocenters. Journal of Nonlinear Science, 8(6), 581–618.

[45] Hauksson, E. (2011). Crustal geophysics and seismicity in southern California. Geophysical Journal International, 186(1), 82–98.

[46] Hauksson, E., Yang, W., & Shearer, P.M. (2013). Waveform relocated earthquake catalog for Southern California, 1981–2013. Retrieved from http://www.data.scec.org/research-tools/downloads.html

[47] Hauksson, E., Yang, W., & Shearer, P.M. (2012). Waveform relocated earthquake catalog for Southern California (1981 to 2011). Bulletin of the Seismological Society of America, 102(5), 2239-2244. doi: 10.1785/0120120010

[48] Helmstetter, A., Kagan, Y.Y., & Jackson, D.D. (2007). High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California. Seismological Research Letters, 78(1), 78–86.

[49] Hicks, A. (2011). Clustering in multidimensional spaces with applications to statistical analysis of earthquake clustering (MSc Thesis). Department of Mathematics and Statistics, University of Nevada, Reno.

[50] Holtkamp, S.G., Pritchard, M.E., & Lohman, R.B. (2011). Earthquake swarms in South America. Geophysical Journal International, 187(1), 128–146.

[51] Jamtveit, B., Ben-Zion, Y., Renard, F., & Austrheim, H. (2018). Earthquake-induced transformation of the lower crust. Nature, 556, 487–491. https://doi.org/10.1038/s41586-018-0045-y.

[52] Joe, H. (2014). Dependence modeling with copulas. CRC press.

[53] Kagan, Y. (2007). Earthquake spatial distribution: The correlation dimension. Geophys. J. Int., 168(3), 1175–1194.

[54] Kagan, Y. (2010). Statistical distributions of earthquake numbers: consequence of branching process. Geophys. J. Int., 180, 1313–1328.

[55] Kagan, Y. Y. (2003). Accuracy of modern global earthquake catalogs. Phys. Earth Planet. Inter., 135(2), 173–209.

[56] Kagan, Y. Y. (2004). Short-term properties of earthquake catalogs and models of earthquake source. Bull. seismol. Soc. Am., 94(4), 1207–1228.

[57] Kato, A., & Ben-Zion, Y. (2021). The generation of large earthquakes. Nat. Rev. Earth Environ. 2, 26–39. doi: 10.1038/s43017- 020-00108-w.

[58] Keilis-Borok, V. I., & Kossobokov, V. G. (1990). Premonitory activation of earthquake flow: Algorithm M8. Phys. Earth Planet. Inter., 61, 73–83.

[59] Kendall, M. G., & Stuart, A. (1979). The advanced theory of statistics (Vol. 2). Macmillan.

[60] Kisslinger, C. (1996). Aftershocks and fault-zone properties. Adv. Geophys., 38, 1–36.

[61] Kisslinger, C., & Jones, L. M. (1991). Properties of aftershock sequences in southern California. J. Geophys. Res., 96, 11,947–11,958.

[62] Kovchegov, Y., Zaliapin, I., & Ben-Zion, Y. (2022). Invariant Galton–Watson branching process for earthquake occurrence. Geophysical Journal International, 231(1), 567-583. https://doi.org/10.1093/gji/ggac204

[63] Kreemer, C., Blewitt, G., & Klein, E. C. (2014). A geodetic plate motion and global strain rate model. Geochem., Geophys., Geosyst., 15(10), 3849–3889.

[64] Kruskal, W., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc., 47(260), 583–621.

[65] Kuge, K. (1992). Systematic difference in the ISC body-wave magnitudeseismic moment relationship between intermediate and deep earthquakes. Bull. seismol. Soc. Am., 82(2), 819–835.

[66] Lengliné, O., Enescu, B., Peng, Z., Shiomi, K. (2012). Decay and expansion of the early aftershock activity following the 2011, MW9.0 Tohoku earthquake. Geophysical Research Letters. 39. 18309-. 10.1029/2012GL052797.

[67] Lockner, D., Walsh, J., & Byerlee, J. (1977). Changes in seismic velocity and attenuation during deformation of granite. Journal of Geophysical Research, 82, 5374-5378.

[68] Lolli, B., & Gasperini, P. (2006). Comparing different models of aftershock rate decay: the role of catalog incompleteness in the first times after main shock. Tectonophysics, 423(1), 43-59.

[69] Luen, B., & Stark, P. B. (2012). Poisson tests of declustered catalogs. Geophysical Journal International. doi:10.1111/j.1365-246X.2012.05400.X.

[70] Lyakhovsky, V., Ben-Zion, Y., & Agnon, A. (1997). Distributed damage, faulting, and friction. Journal of Geophysical Research: Solid Earth, 102, 27635-27649.

[71] Marzocchi, W., & Sandri, L. (2003). A review and new insights on the estimation of the b-value and its uncertainty. Annals of Geophysics, 46(6), 1271-1282.

[72] Mavko, G., Mukerji, T., & Dvorkin, J. (2009). The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media. Cambridge University Press.

[73] McKenzie, E. D. (1982). Product autoregression: A time-series characterization of the gamma distribution. Journal of Applied Probability, 19, 463-468.

[74] Mogi, K. (1969). Some features of recent seismic activity in and near Japan 2: Activity before and after great earthquakes. Bulletin of the Earthquake Research Institute, University of Tokyo, 47, 395-417.

[75] Molchan, G., & Kronrod, T. (2009). The fractal description of seismicity. Geophysical Journal International, 179(3), 1787-1799.

[76] Moradpour, J., Hainzl, S., & Davidsen, J. (2014). Nontrivial decay of aftershock density with distance in Southern California. Journal of Geophysical Research: Solid Earth, 119, 5518-5535.

[77] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24(3), 69-71.

[78] Narteau, C., Shebalin, P., & Holschneider, M. (2002). Temporal limits of the power law aftershock decay rate. Journal of Geophysical Research: Solid Earth, 107(B12), 2359. doi:10.1029/2002JB001868.

[79] Nelsen, R. B. (2006). An introduction to copulas. Springer Science & Business Media.

[80] Ogata, Y. (1985). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes [Technical report]. Research Memo., No. 288, Inst. Statist. Math., Tokyo.

[81] Ogata, Y. (1988). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. J. Amer. Statist. Assoc., 83, 9–27.

[82] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. Annals of the Institute of Statistical Mathematics, 50(2), 379-402.

[83] Ogata, Y. (1999). Seismicity analysis through point-process modeling: a review. In M. Wyss, K. Shimazaki, & A. Ito (Eds.), Seismicity Patterns, their Statistical Significance and Physical Meaning (pp. 471–507). doi.org/10.1007/978-3-0348-8677-214.

[84] Omori, F. (1894). On the aftershocks of earthquakes. J. College Sci., Imperial Univ. Tokyo, 7, 111–200.

[85] Osher, S., & Fedkiw, R. (2003). Level set methods and dynamic implicit surfaces. Springer.

[86] Peng, S., & Johnson, A. M. (1972). Crack growth and faulting in cylindrical specimens of Chelmsford granite. Int. J. Rock Mech. Min. Sci., 9, 37–86.

[87] Peng, Z., Vidale, J. E., Ishii, M., & Helmstetter, A. (2007). Seismicity rate immediately before and after main shock rupture from high-frequency waveforms in Japan. J. geophys. Res.: Solid Earth, 112, B03306. doi:10.1029/2006JB004386

[88] Renard, F., Weiss, J., Mathiesen, J., Ben Zion, Y., Kandula, N., & Cordonnier, B. (2018). Critical evolution of damage towards system-size failure in crystalline rock. J. Geophys. Res., 123, 1969–1986. https://doi.org/10.1002/2017JB014964.

[89] Reverso, T., Marsan, D., & Helmstetter, A. (2015). Detection and characterization of transient forcing episodes affecting earthquake activity in the Aleutian Arc system. Earth planet. Sci. Lett., 412, 25–34.

[90] Richards-Dinger, K. B., & Shearer, P. M. (2000). Earthquake locations in southern California obtained using source-specific station terms. J. geophys. Res.: Solid Earth (1978–2012), 105(B5), 10 939–10 960.

[91] Rohm, A. H. E., Trampert, J., Paulssen, H., & Snieder, R. K. (1999). Bias in reported seismic arrival times deduced from the ISC Bulletin. Geophys. J. Int., 137(1), 163–174.

[92] Ross, Z. E., Ben-Zion, Y., White, M. C., & Vernon, F. L. (2016). Analysis of earthquake body wave spectra for potency and magnitude values: implications for magnitude scaling relations. Geophys. J. Int., 207, 1158–1164. https://doi.org/10.1093/gji/ggw327

[93] Ross, Z. E., Rollins, C., Cochran, E. S., Hauksson, E., Avouac, J.-P., & Ben-Zion, Y. (2017). Aftershocks driven by afterslip and fluid pressure sweeping through a fault-fracture mesh. Geophysical Research Letters, 44, 8260–8267. https://doi.org/10.1002/2017GL074634

[94] Saichev, A., Helmstetter, A., & Sornette, D. (2005). Power-law distributions of offspring and generation numbers in branching models of earthquake triggering. Pure and Applied Geophysics, 162, 1113–1134.

[95] Scholz, C. H. (2002). The mechanics of earthquakes and faulting. Cambridge University Press.

[96] Shcherbakov, R., & Turcotte, D. L. (2004). A modified form of Bath's law. Bulletin of the Seismological Society of America, 94, 1968–1975.

[97] Shearer, P. M. (2012). Self-similar earthquake triggering, Bath's law, and foreshock/aftershock magnitudes: Simulations, theory, and results for southern California. Journal of Geophysical Research, 117, B06310.

[98] Shimazaki, K. (1971). On periodicity of earthquake occurrences. Kagaku, 41, 688–689. (in Japanese)

[99] Schoenball, M., Davatzes, N. C., & Glen, J. M. G. (2015). Differentiating induced and natural seismicity using space-time-magnitude statistics applied to the Coso Geothermal field. Geophysical Research Letters, 42. https://doi.org/10.1002/2015GL064772

[100] Schorlemmer, D., & Woessner, J. (2008). Probability of detecting an earthquake. Seismological Society of America, 98(5), 2103–2117.

[101] Storchak, D. A., Bird, A. L., & Adams, R. D. (2000). Discrepancies in earthquake location between ISC and other agencies. Journal of Seismology, 4(3), 321–331.

[102] Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of Graduate Medical Education, 4(3), 279–282.

[103] Tamaribuchi, K., Yagi, Y., Enescu, B., & Hirano, S. (2018). Characteristics of foreshock activity inferred from the JMA earthquake catalog. Earth, Planets and Space, 70, 90. https://doi.org/10.1186/s40623-018-0869-2

[104] Taylor, M. E. (2018). Measure theory and integration. American Mathematical Society.

[105] Tinti, S., & Mulargia, F. (1987). Confidence intervals of b values for grouped magnitudes. Bulletin of the Seismological Society of America, 77(6), 2125–2134.

[106] Utsu, T. (1961). A statistical study on the occurrence of aftershocks. Geophysical Magazine, 30, 521–605.

[107] Utsu, T. (1975). Correlation between shallow earthquakes in Kwanto Region and intermediate earthquakes in Hida Region, Central Japan. Zisin (J. Seismol. Soc., Japan) Ser. 2 (28), 303–311. (in Japanese)

[108] Utsu, T., Ogata, Y., & Matsu'ura, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity. Journal of Physics of the Earth, 43, 1–33.

[109] Utsu, T., & Ogata, Y. (1997). Statistical analysis of seismicity. In Algorithms for Earthquake Statistics and Prediction, IASPEI Software Library 6, 13-94, International Association of Seismology and Physics of the Earth's Interior in collaboration with the Seismological Society of America.

[110] Utsu, T. (2002). Statistical features of seismology. In International Handbook of Earthquake and Engineering Seismology, Part A (pp. 719–732). Academic Press.

[111] van der Elst, N. J. (2017). Accounting for orphaned aftershocks in the earthquake background rate. Geophysical Journal International, 211(2), 1108–1118.

[112] Veen, A., & Schoenberg, F. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. Journal of the American Statistical Association, 103, 614–624.

[113] Vere-Jones, D., & Ozaki, T. (1982). Some examples of statistical estimation applied to earthquake data, 1. cyclic Poisson and self-exciting models. Annals of the Institute of Statistical Mathematics, 34B, 189–207.

[114] Vere-Jones, D. (1985). The detection and estimation of periodicities in point process data. Technical report, ISOR, Victoria Univ. of Wellington.

[115] Vidale, J. E., Boyle, K. L., & Shearer, P. M. (2006). Crustal earthquake bursts in California and Japan: their patterns and relation to volcanoes. Geophysical Research Letters, 33, L20313. https://doi.org/10.1029/2006GL027723

[116] Vidale, J. E., & Shearer, P. M. (2006). A survey of 71 earthquake bursts across southern California: exploring the role of pore fluid pressure fluctuations and aseismic slip as drivers. Journal of Geophysical Research, 111, B05312. https://doi.org/10.1029/2005JB004034

[117] Villani, C. (2003). Topics in Optimal Transportation. American Mathematical Society.

[118] Villani, C. (2009). Optimal transport: Old and new (Vol. 338). Springer Science & Business Media.

[119] Wang, Q., F. P. Schoenberg, & D. Jackson (2010). Standard errors of parameter estimates in the ETAS model. Bulletin of the Seismological Society of America, 100(5A), 1989–2001.

[120] Wang, Q., D. D. Jackson, & J. Zhuang (2010). Missing links in earthquake clustering models. Geophysical Research Letters, 37(21), doi: 10.1029/2010GL044858.

[121] Wiemer, S., & Katsumata, K. (1999). Spatial variability of seismicity parameters in aftershock zones. Journal of Geophysical Research: Solid Earth (1978-2012), 104(B6), 13,135–13,151.

[122] Wu, C., Meng, X., Peng, Z., & Ben-Zion, Y. (2014). Lack of spatiotemporal localization of foreshocks before the 1999 Mw 7.1 Düzce, Turkey, earthquake. Bulletin of the Seismological Society of America, 104, 560-566.

[123] Zaliapin, I., A. Gabrielov, V. Keilis-Borok, and H. Wong (2008). Clustering analysis of seismicity and aftershock identification, Phys. Rev. Lett. 101, no. 1, 018501, doi: 10.1103/ PhysRevLett.101.018501

[124] Zaliapin, I., & Ben-Zion, Y. (2011). Asymmetric distribution of aftershocks on large faults in California. Geophysical Journal International, 185(3), 1288–1304.

[125] Zaliapin, I., & Ben-Zion, Y. (2013a). Earthquake clusters in southern California I: Identification and stability. Journal of Geophysical Research: Solid Earth, 118(6), 2847-2864.

[126] Zaliapin, I., & Ben-Zion, Y. (2013b). Earthquake clusters in southern California II: Classification and relation to physical properties of the crust. Journal of Geophysical Research, 118(6), 2865–2877.

[127] Zaliapin, I., & Ben-Zion, Y. (2015). Artifacts of earthquake location errors and short-term incompleteness on seismicity clusters in southern California. Geophysical Journal International, 202, 1949–1968.

[128] Zaliapin, I., & Ben-Zion, Y. (2016a). Discriminating characteristics of tectonic and human-induced seismicity. Bulletin of the Seismological Society of America, 106(2), 635-647.

[129] Zaliapin, I., & Ben-Zion, Y. (2016b). A global classification and characterization of earthquake clusters. Geophysical Journal International, 207(1), 608-634.

[130] Zaliapin, I., & Ben-Zion, Y. (2020). Earthquake declustering using the nearest-neighbor approach in space-time-magnitude domain. Journal of Geophysical Research: Solid Earth, e53991. https://doi.org/10.1029/2018JB017120

[131] Zaliapin, I., & Ben-Zion, Y. (2021). Perspectives on clustering and declustering of earthquakes. Seismological Research Letters, 92(1), 386-401. https://doi.org/10.1785/0220210127

[132] Zhang, Q., & Shearer, P. M. (2016). A new method to identify earthquake swarms applied

A Appendix A - Figures of ROC Analysis

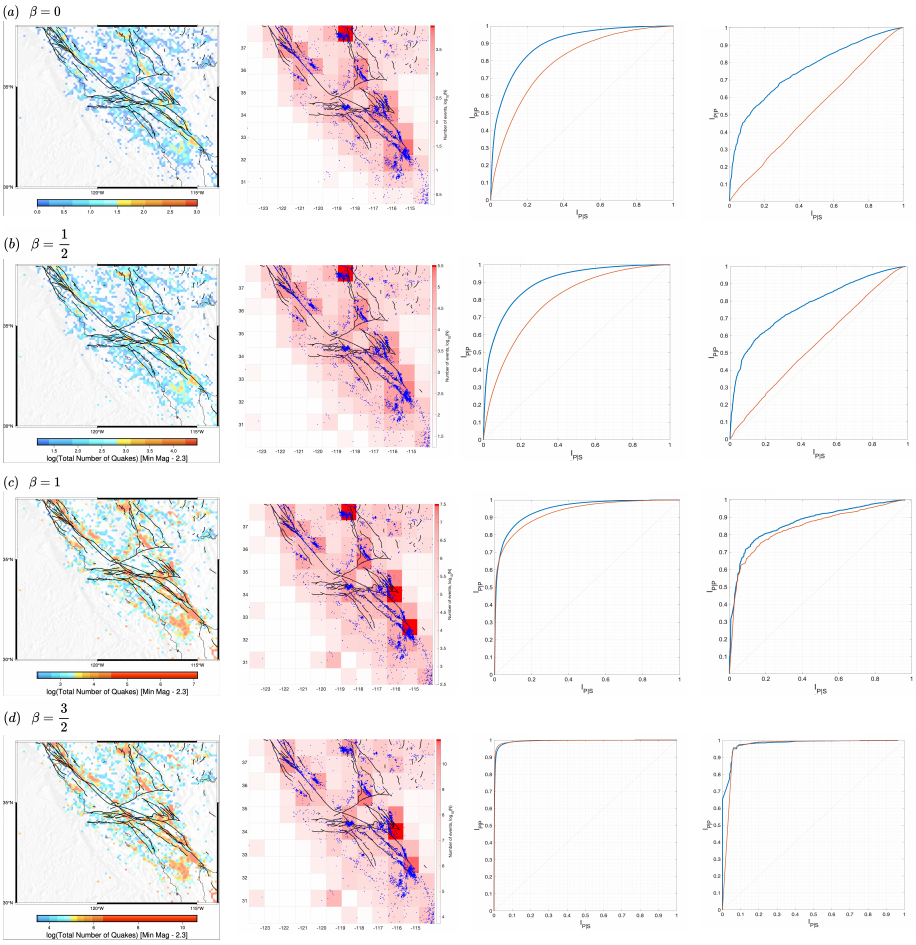

Figure 25: ROC analysis for southern California. Magnitude of completeness 2.3.

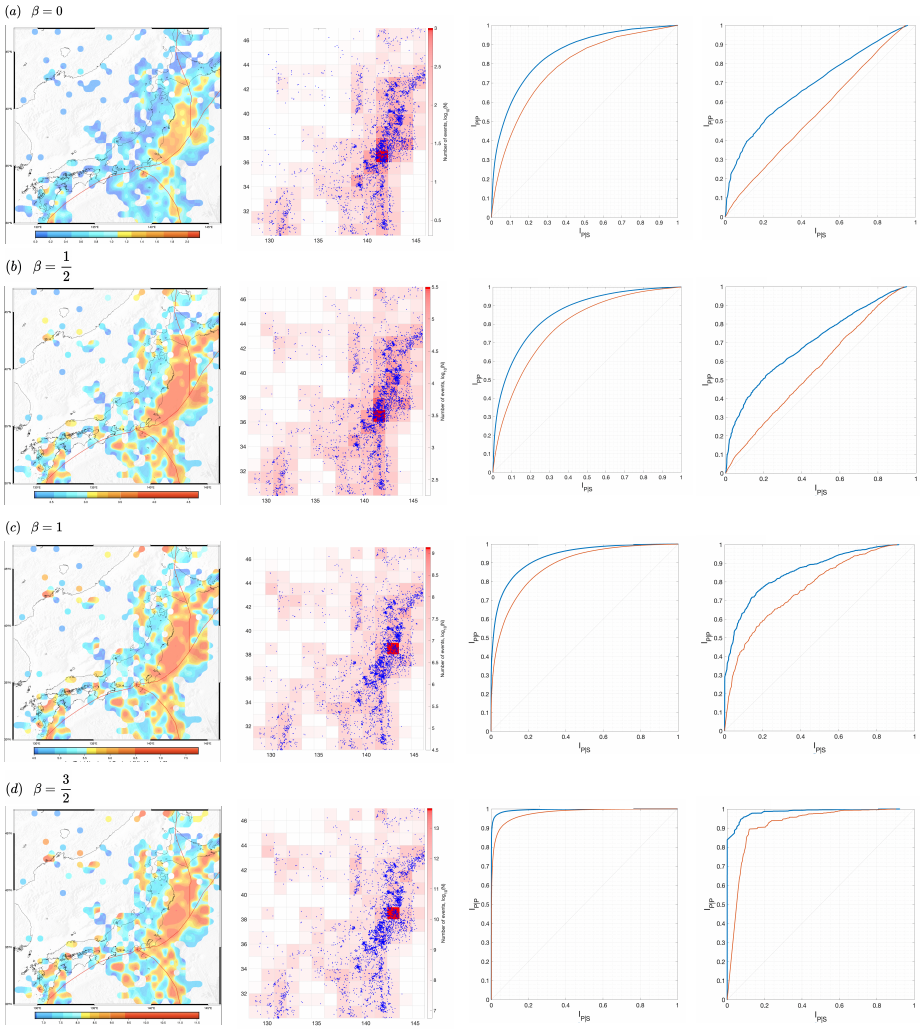

Figure 26: ROC analysis for Japan. Magnitude of completeness 4.5. Colors represent number of quakes on the log scale.

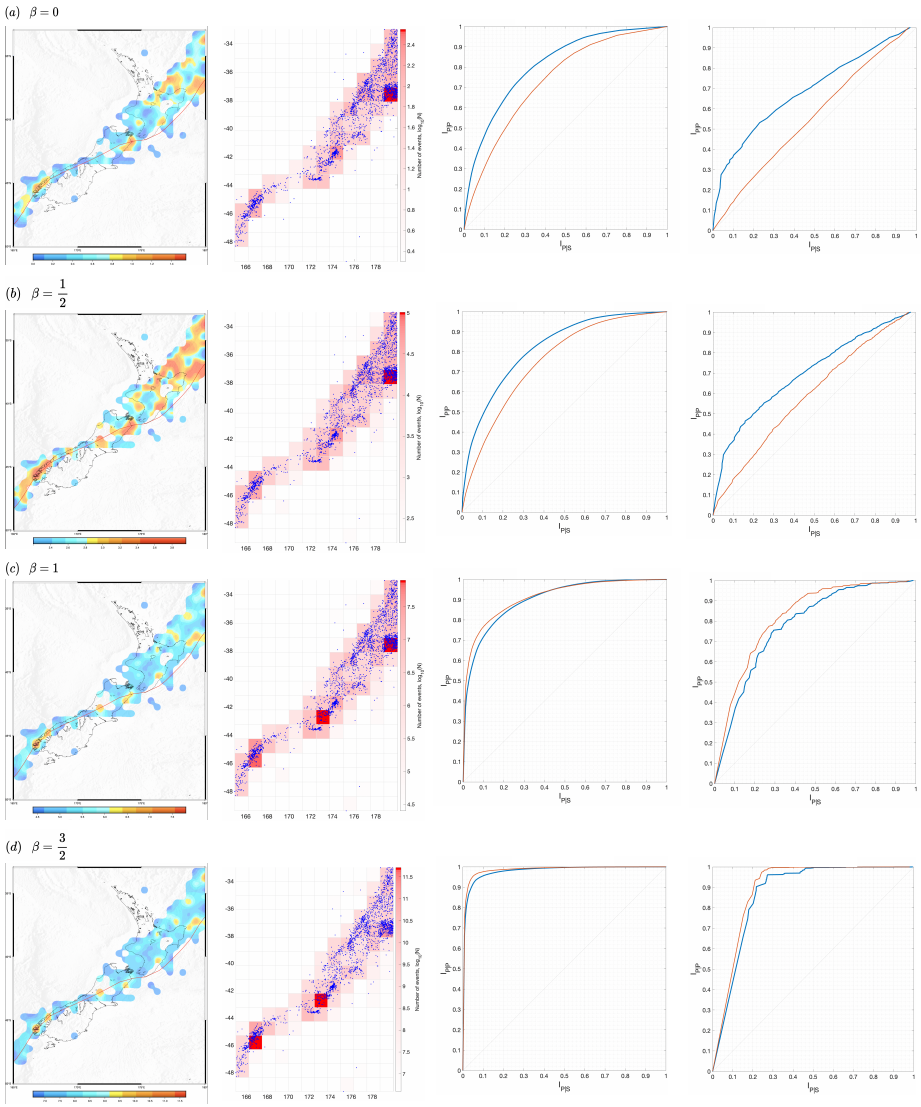

Figure 27: ROC analysis for New Zealand. Magnitude of completeness 4.4. Colors represent number of quakes on the log scale.

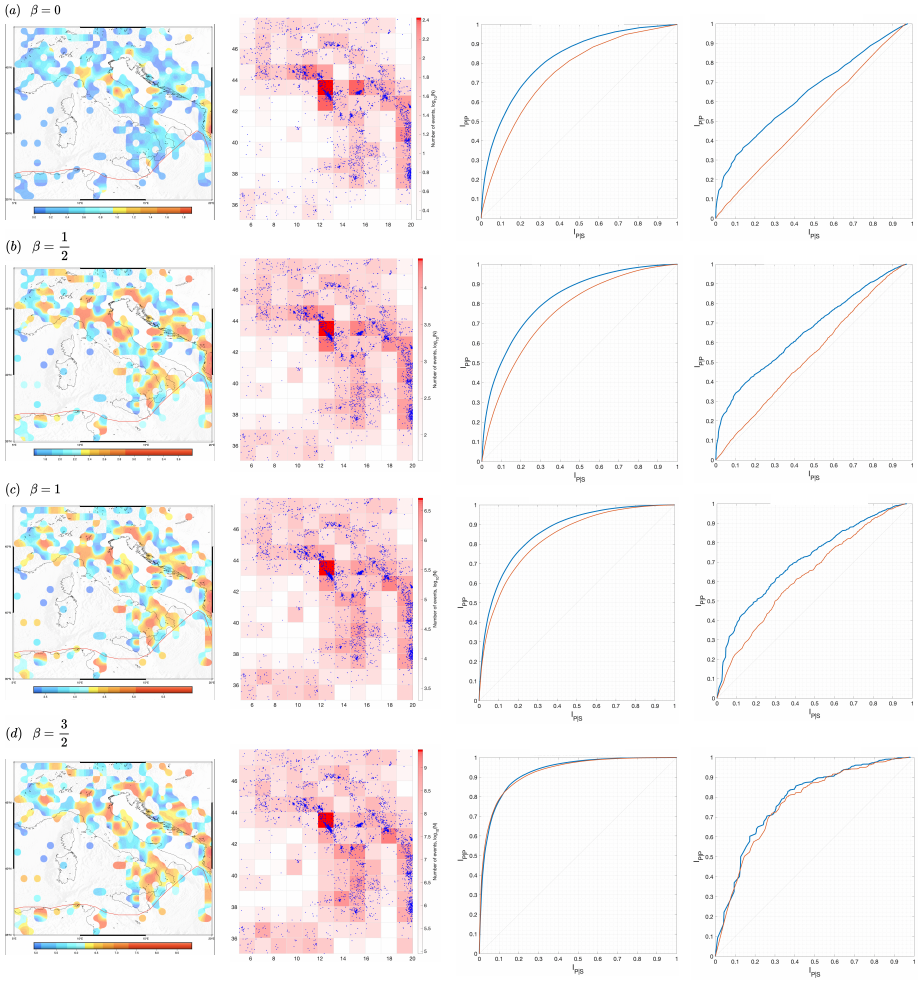

Figure 28: ROC analysis for Italy. Magnitude of completeness 3.3. Colors represent number of quakes on the log scale.

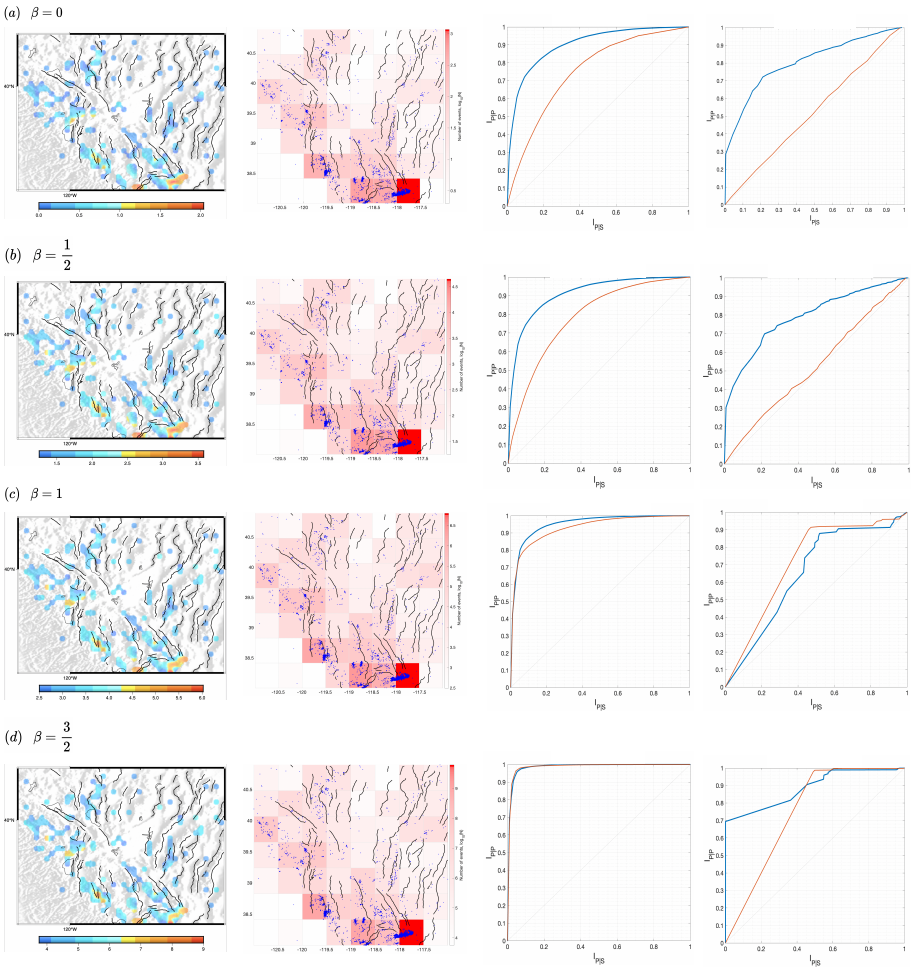

Figure 29: ROC analysis for Reno. Magnitude of completeness 2.5. Colors represent number of quakes on the log scale.